UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

IVO GABRIEL DE ABREU NICOLAIO

DETECÇÃO DE ATAQUES EM DUAS FASES USANDO APRENDIZADO DE MÁQUINAS EM SISTEMAS DE CONTROLE INDUSTRIAL DE INFRAESTRUTURAS CRÍTICAS

CURITIBA

2024

IVO GABRIEL DE ABREU NICOLAIO

DETECÇÃO DE ATAQUES EM DUAS FASES USANDO APRENDIZADO DE MÁQUINAS EM SISTEMAS DE CONTROLE INDUSTRIAL DE INFRAESTRUTURAS CRÍTICAS

Two-phase attack detection using Machine Learning in Critical Infrastructure Industrial Control Systems

Dissertação apresentado como requisito para obtenção do título de “Mestre em Ciências” - Área de Concentração: Telecomunicações e Redes do Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná.

Orientador: Profª. Drª. Anelise Munaretto Fonseca

Coorientador: Prof. Dr. Mauro Sérgio Pereira Fonseca

CURITIBA

2024

4.0 Internacional

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

IVO GABRIEL DE ABREU NICOLAIO

DETECÇÃO DE ATAQUES EM DUAS FASES USANDO APRENDIZADO DE MÁQUINAS EM SISTEMAS DE CONTROLE INDUSTRIAL DE INFRAESTRUTURAS CRÍTICAS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Ciências da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Telecomunicações E Redes.

Data de aprovação: 08 de Março de 2024

Dr. Mauro Sergio Pereira Fonseca, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Michele Nogueira Lima, Doutorado - Universidade Federal de Minas Gerais (Ufmg)

Dr. Paulo Cesar Pellanda, Doutorado - Instituto Militar de Engenharia (Ime)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 09/03/2024.

AGRADECIMENTOS

Agradeço a Deus por tantas graças concedidas, a despeito das minhas fraquezas.

Agradeço aos meus orientadores, Profª. Drª. Anelise Munaretto e Prof. Dr. Mauro Fonseca, por me guiarem com paciência desde o começo desta trajetória, aconselhando e direcionando os esforços para os melhores resultados e dispondo de tempo também para assuntos pessoais do dia a dia.

Agradeço ao 11º Centro de Telemática por disponibilizar a infraestrutura em seu *data-center* para execução dos extensos algoritmos e tratamento do grande volume de dados.

Expresso minha gratidão aos Professores Valmir de Oliveira, Thiago Silva, Jean Carlos e André Lazzaretti pelo entusiasmo e qualidade das aulas ministradas.

Registro a minha satisfação com a Secretaria e Coordenação do Programa, sempre atendendo as solicitações de forma ágil e eficaz.

Agradeço ao meus amigos, que acreditavam mais que eu em dias melhores antes de eu conseguir vir para Curitiba cursar o mestrado.

“Eu vos disse tais coisas para terdes paz em mim. No mundo tereis tribulações, mas tende coragem: eu venci o mundo!"
João 16:33

RESUMO

Com o advento da Indústria 4.0, Sistemas de Controle Industrial (ICS) estão cada vez mais integrados às redes corporativas e à *internet*. Se, por um lado, há a vantagem do acesso à informação em tempo real e execução remota de rotinas do processo industrial, por outro, há o aumento da superfície de ataque cibernético com motivações diversas. Tais ambientes não são exclusivos de indústrias de corporações, mas são parte essencial de infraestruturas críticas de diversos setores, como energia, água e nuclear. Ataques em tais ambientes trazem impactos se não difíceis, impossíveis de calcular, e, não raro, têm sido abordados por matérias jornalísticas, como no caso da guerra entre Rússia e Ucrânia. Estratégias de segurança já são empregadas no ambiente de Tecnologia da Informação (TI), porém não são eficazes na detecção de ataques no ambiente de Tecnologia da Operação (TO). Este trabalho propõe uma composição de aprendizado supervisionado e não supervisionado baseada em classificadores tradicionais de Aprendizado de Máquinas para a detecção de ameaças, causadas especialmente por ataques de injeção de dados falsos, em dados de um *dataset* de rede industrial. A proposta é aplicada em *dataset* proveniente de simulação em *Hardware-in-the-Loop* integrado com componentes reais, representativo de uma usina de geração de energia elétrica que sofre ataques de injeção e manipulação dos dados trafegados pelo ICS. São criados vários cenários derivados do *dataset* inicial e as detecções avaliadas pelas métricas acurácia, *recall*, AUC e F_1-*score*. Os resultados da composição são comparados com a detecção por assinatura, obtendo ganhos médios relativos variando entre 32,3% e 179,15%, conforme a métrica empregada. Foi possível obter, com a composição, valores absolutos de métricas superiores a 0,952 e reduções de taxa de falsos negativos na média de até 19,36%, conforme a métrica. A análise de falsos positivos e falsos negativos abriu espaço para escolha de diferentes métricas para avaliação da composição, conforme políticas da organização responsável pela infraestrutura crítica.

Palavras-chave: infraestrutura crítica; segurança cibernética; sistemas de controle industrial; aprendizado de máquinas; detecção de anomalias.

ABSTRACT

The advent of Industry 4.0 has been led an increasingly integration of Industrial Control Systems (ICS) into corporate networks and to the internet. If there is the advantage of real time access to processes information and remote execution of industrial process routines, on the other hand there is an increase on the cyberattack surface, backed by different motivations. Such environments are not exclusive to corporate industries, but are an essential part of critical infrastructure from various sectors, like energy, water and nuclear plants. Attacks in those environments have impacts that are difficult or even impossible to measure, and have often been covered by journalistic articles, as in the case of the war between Russia and Ukraine. Security strategies are already used in the Information Technology (IT) networks, but they are not effective in detecting attacks in the Operation Technology (OT) environment. This work proposes a composition of supervised and unsupervised learning based on traditional Machine Learning classifiers for detecting threats, mainly as a result of False Data Injection attacks, in data from an industrial network dataset. The proposal is applied to a dataset from a Hardware-in-the-Loop simulation integrated with real components, representative of a Smart Grid that suffers injection attacks and manipulation of data transmitted by the ICS components. Several scenarios were derived from the initial dataset and the detections were evaluated by the metrics accuracy, recall, AUC and F1-score. The composition results are compared with signature detection, obtaining average relative gains between 32.3% and 179.15% depending on the metric used. The composition allowed absolute values of metrics greater than 0.952 and false negative rates reductions up to 19.36% on average depending on the metric. The analysis of false positives and false negatives made way for different choices of metric evaluations to apply on the composition, in accordance with the polices of the organization responsible for the critical infrastructure.

Keywords: critical infrastructure; cybersecurity; industrial control systems; machine learning; anomaly detection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

AUC	*Area Under the Receiver Operating Characteristic Curve*
CPS	Sistemas ciberfísicos, do inglês *Cyber Physical System*
CSOC	Centro de Operações de Segurança Cibernética, do inglês *Cyber Security Operations Centre*
CSV	*Comma-separated values*
DNS	Sistema de Nomes de Domínio, do inglês *Domain Name System*
END	Estratégia Nacional de Defesa
FDI	Injeção de Dados Falsos, do inglês *False Data Injection*
FN	Falsos Negativos
FNR	Taxa de falsos negativos, do inglês *False Negative Rate*
FP	Falsos Positivos
FPR	Taxa de falsos positivos, do inglês *False Positive Rate*
GSI/PR	Gabinete de Segurança Institucional da Presidência da República
HAI	*HIL-based Augmented ICS Security Dataset*
HIL	*Hardware-in-the-Loop*
HMI	Interface Homem-Máquina, do inglês *Human-Machine Interface*
ICS	Sistema de Controle Industrial, do inglês *Industrial Control System*
IDS	Sistema de Detecção de Intrusão, do inglês *Intrusion Detection System*
IIoT	*Industrial Internet of Things*
IP	Protocolo de *Internet*, do inglês *Internet Protocol*
KNN	*K-Nearest Neighbors*
LASC	Laboratório de Simulação Cibernética
LASEC²	Laboratório de Segurança Eletrônica, de Comunicações e Cibernética

OCSVM	*One-Class Support Vector Machine*
OD	*Outlier Detection*
OND	Objetivo Nacional de Defesa
P1	Processo da Caldeira ou *Boiler Process* do HAI
P2	Processo da Turbina ou *Turbine Process* do HAI
P3	Processo da hidroelétrica ou *Water-treatment Process* do HAI
P4	Processo do simulador HIL ou *HIL Simulation Process* do HAI
PERA	*Purdue Enterprise Reference Architecture*
PLC	Controladores Lógicos Programáveis, do inglês *Programmable Logic Controller*
PNSIC	Política Nacional de Segurança de Infraestruturas Críticas
RFECV	*Recursive Feature Elimination with Cross-Validation*
RTU	Unidades Terminais Remotas, do inglês *Remote Terminal Unit*
SCADA	*Supervisory Control and Data Acquisition*
SMOTE	*Synthetic Minority Over-sampling Technique*
SQL	Linguagem de Consulta Estruturada, do inglês *Structured Query Language*
TCP	Protocolo de Controle de Transmissão, do inglês *Transmission Control Protocol*
TI	Tecnologia da Informação
TO	Tecnologia da Operação
UDP	Protocolo de Datagrama do Usuário, do inglês *User Datagram Protocol*
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
VQ	*Vector Quantization*

SUMÁRIO

1 INTRODUÇÃO

Em 24 de fevereiro de 2022 a Rússia invadiu a Ucrânia, desencadeando o maior conflito armado na Europa desde a 2ª Guerra Mundial (PLOKHY, 2023). Segundo o Centro de Segurança Cibernético Canadense (2022), o movimento, apesar de surpreender a comunidade global, foi precedido de diversos alertas midiáticos, mobilizações militares e eventos cibernéticos de grande impacto nas infraestruturas críticas, principalmente do setor de energia (CANADA, 2022).

O ataque cibernético a infraestruturas críticas não está restrito a países em conflitos armados. Atores isolados motivados por extorsão financeira, obtenção de dados sigilosos, roubo de patentes, ou para aumentar a fama nos grupos ciberativistas podem atuar para interromper serviços essenciais como o fornecimento de água e gás. Os danos resultantes de ataques em infraestruturas críticas não se limitam a prejuízos financeiros, podem afetar o meio ambiente, a ordem social e, no pior caso, resultar em perdas de vidas humanas (ANI; HE; TIWARI, 2016).

No primeiro semestre de 2023, foram bloqueados artefatos maliciosos em 34% de computadores em ambientes industriais monitorados pela Kaspersky, tendo a *internet* a primeira posição das fontes de ameaças. Nas categorias de setores industriais, do segundo semestre de 2022 para 2023, enquanto houve uma redução de 4,7% de bloqueios no setor de óleo e gás, no setor de energia houve um aumento de 1,5%, alcançando 36% dos bloqueios (KASPERSKY, 2023).

Apesar de contar com reconhecidos mecanismos de segurança (e.g. *firewalls*, *softwares* antivírus, sistemas de detecção a intrusão, *honeypots*), tais medidas são limitadas à ação nos ambientes de Tecnologia da Informação (TI), sendo pouco eficientes quando o vetor de exploração e ataque atua além dessa rede. Nesse sentido, infraestruturas críticas dos setores de energia, água, nuclear e químico, por exemplo, possuem verdadeiras plantas industriais para provimento de serviço essencial, integradas a uma rede corporativa, cada vez mais distribuída geograficamente (ALANAZI; MAHMOOD; CHOWDHURY, 2023).

Portanto, além da atuação na rede corporativa, o emprego de estratégias de segurança na rede industrial torna-se essencial para combater as ameaças cibernéticas, ainda mais quando considerado o advento da Indústria 4.0, com o aumento da interconexão de dispositivos de campo à *internet* e consequente aumento da superfície de ataque por agentes externos.

1.1 Considerações iniciais

Tendo em vista o contexto da necessidade de atuação nas redes industriais de infraestruturas críticas, este trabalho focou na detecção de ataques cibernéticos em dados de uma rede industrial típica de um ambiente de tecnologia da operação, vislumbrando a complementação com os esforços na rede TI, mais amadurecidos e com diversas pesquisas e publicações.

No Paraná, a principal infraestrutura crítica do setor de energia é a Binacional Itaipu, em Foz do Iguaçu, que em conjunto a Fundação Parque Tecnológico Itaipu e o Comando do Exército assinaram um Acordo de Cooperação Técnico para o investimento em pesquisa e criação do Centro de Estudos Avançados em Proteção de Estruturas Estratégicas. No futuro é possível que sejam disponibilizados dados de uma rede industrial baseada nos equipamentos presentes na usina, para estudo de detecção de ataques, anomalias e efeitos de *malwares*.

Atualmente, dentre os *datasets* de domínio público disponíveis para pesquisa em sistemas industriais, foi verificado um recente que se sobressai em relação aos demais pelo uso de simulação em *hardware* e integração com componentes reais (SHIN *et al.*, 2020). O *dataset* foi criado para simular os efeitos de ataques em componentes ciberfísicos de uma usina de geração de energia elétrica, observando as alterações sistêmicas nas medições de sensores e controle dos atuadores do ambiente de campo. Apesar de classificados por alguns autores como ataques do tipo Injeção de Dados Falsos, do inglês *False Data Injection* (FDI), o termo pode levar a entendimentos equivocados ao assumir que dados falsos são introduzidos na sequência temporal de leitura dos equipamentos. Trata-se, de fato, na alteração dos dados gerados pelos sensores antes de darem entrada nas próximas etapas do processo industrial. A injeção de dados falsos, nesses casos, sobrescreve os dados reais e podem ter objetivos específicos (alteração de variáveis específicas para valores previamente calculados) ou em busca de aproveitamento do êxito da manipulação desses dados qualquer distorção que possa resultar em danos na infraestrutura crítica causados por essas alterações, sem considerar os métodos pelos quais foram produzidas.

Este trabalho propõe, em comparação aos trabalhos de Xingchao (2020), Mokhtari *et al.* (2021) e Wang *et al.* (2023), uma abordagem baseada na composição de classificadores supervisionados e não supervisionados tradicionais, a avaliação do desempenho de detecção sob as métricas F_1*-score*, *recall*, *Area Under the Receiver Operating Characteristic Curve* (AUC) e acurácia e a comparação da detecção composta com a detecção por assinatura.

1.2 Objetivos

Este trabalho tem como objetivo realizar a detecção de anomalias nos dados de uma simulação de rede industrial de infraestrutura crítica do setor de energia, causadas por ataques cibernéticos que atuem nos componentes ciberfísicos, utilizando técnicas de Aprendizado de Máquinas, em especial na tarefa de classificação na composição das abordagens supervisionadas e não supervisionadas, configurando uma detecção em duas fases. Apesar do foco nas alterações dos dados gerados pelos equipamentos industriais causadas com a intenção de ataques, tais anomalias podem ser decorrentes de outras ameaças, decorrentes, por exemplo, da presença humana nesses ambientes e suas interações, ainda que acidentais, com componentes mecânicos da planta industrial.

Em específico, objetiva-se a avaliação do desempenho por métricas além da majoritariamente utilizada (acurácia), comparando os resultados da detecção pela composição com a detecção por assinatura, e verificando a viabilidade de adoção dessa abordagem.

1.3 Justificativa

Na literatura de segurança cibernética de Sistemas de Controle Industrial (ICS) de infraestruturas críticas, foi observado o uso majoritário de técnicas de *Deep Learning* para a detecção da alteração de dados causados por intrusão, *malwares* e demais atividades maliciosas, em geral justificando o uso pelos bons resultados na métrica acurácia.

Após revisar os trabalhos relacionados ao *dataset* empregado, foi verificada uma lacuna quanto ao uso de métodos tradicionais de aprendizado de máquina, com poucos trabalhos aplicando a classificação supervisionada por meio de algoritmos mais reconhecidos e verificando o desempenho pela métrica acurácia.

Apesar de poder apresentar resultados superiores a 0,9, a métrica acurácia pode levar a interpretações equivocadas do modelo quando utilizada com problemas de classes desbalanceadas. Ademais, ainda que o aprendizado supervisionado seja eficaz para detecções, ele não é adequado para ataques inéditos ou anomalias desconhecidas pelo modelo aprendido. A abordagem não supervisionada atua nesse cenário, reconhecendo padrões intrínsecos dos dados mesmo desconhecendo a classe a que cada instância faz parte.

Alguns trabalhos da literatura propõem composições de técnicas do mesmo grupo ou de um grupo e de outro, porém, conforme discutido no Capítulo 3, não foi verificado o uso conjunto de algoritmos tradicionais supervisionados e não supervisionados para detecção de ataques em ICS da forma proposta neste trabalho.

1.4 Estrutura do trabalho

Este trabalho está estruturado na intenção da construção da linha de raciocínio que levaram à proposta da pesquisa, sua execução e resultados. Nesse sentido, no Capítulo 2, são apresentados os conceitos e a literatura de amparo para a contextualização do problema. No Capítulo 3, o escopo é aprofundado e são apresentados os trabalhos que contribuem para a pesquisa em detecção de ataques em ICS e levaram à identificação da lacuna de atuação. O Capítulo 4 apresenta as ferramentas, ambiente de execução dos experimentos e a metodologia para alcançar a proposta. O Capítulo 5 apresenta os resultados obtidos com a aplicação da proposta, estatísticas e *insights* para a detecção de ataques. Por fim, o Capítulo 6 traz as observações finais do trabalho, encerrando com as referências utilizadas.

2 REFERENCIAL TEÓRICO

Neste capítulo, são apresentados os conceitos, amparos e revisão da literatura que dão base à pesquisa realizada. Visa nivelar entendimentos e reduzir ambiguidades conceituais. São tratados os tópicos de infraestruturas críticas, redes industriais e ICS, aspectos de segurança cibernética e Aprendizado de Máquinas.

2.1 Infraestruturas Críticas

O termo Infraestruturas Críticas vem crescendo em importância com o advento da Indústria 4.0 e da Internet das Coisas, principalmente pelas preocupações concernentes aos impactos causados por ataques cibernéticos, não limitados a prejuízos materiais (AZAMBUJA; ALMEIDA, 2021). A Política Nacional de Segurança de Infraestruturas Críticas (PNSIC) conceitua o termo como “instalações, serviços, bens e sistemas cuja interrupção ou destruição, total ou parcial, provoque sério impacto social, ambiental, econômico, político, internacional ou à segurança do Estado e da sociedade” (BRASIL, 2018).

As ameaças a essas estruturas variam entre desastres naturais, acidentes, calamidades, sabotagens, espionagem militar e ataques cibernéticos. Exemplos de ataques cibernéticos como o citado no Capítulo 1 estão promovendo não apenas a pesquisa acadêmica nos esforços de proteção mas também o desenvolvimento de amparo legal e da conscientização do problema. Nos Estados Unidos a *Cybersecurity and Infraestructure Security Agency* foi criada em 2018 com a missão de liderar os esforços nacionais na gestão da segurança das infraestruturas críticas frente às ameaças não apenas de *hackers* casuais mas principalmente de grupos suportados por estados-nação (EUA, 2018). Medidas semelhantes estão sendo adotadas por países da União Europeia, como o Reino Unido (REINO UNIDO, 2023) e França (FRANÇA, 2018), bem como pela Rússia (PURSIAINEN, 2020).

No Brasil, o Gabinete de Segurança Institucional da Presidência da República (GSI/PR) é o órgão responsável pela implementação e gestão do Sistema Integrado de Dados de Segurança de Infraestruturas Críticas, que é um dos instrumentos da PNSIC, além da Estratégia Nacional de Segurança de Infraestruturas Críticas (BRASIL, 2018). Antes da edição do Decreto nº 9.573, o GSI/PR já conduzia ações para propiciar o tratamento adequado ao tema, com a criação de grupos de trabalho para identificar e classificar essas estruturas e avaliar medidas preventivas contra calamidades, bem como foram estabelecidas cinco áreas prioritárias no esforço de segurança de infraestruturas críticas: energia, transporte, água, telecomunicações e sistemas financeiros (BRASIL, 2008).

Além da preocupação com as infraestruturas críticas, houve o entendimento da gênese do conceito de espaço cibernético, o qual agrega ameaças silenciosas e de elevados impactos nessas estruturas, de tal modo que também em 2008 a Estratégia Nacional de Defesa (END) estabeleceu três setores prioritários para a Defesa Nacional, reforçados até a última edição

disponível: nuclear, cibernético e espacial (BRASIL, 2016). Em confluência com a END o Livro Branco de Defesa Nacional confirmou ao Exército Brasileiro a atribuição da coordenação estratégica do setor cibernético, destacando o risco às infraestruturas críticas, impactos à soberania nacional e necessidade de capacitação e pesquisa científica (BRASIL, 2012).

Na END consta como Objetivo Nacional de Defesa (OND) número um, dentre oito, "garantir a soberania, o patrimônio nacional e a integridade territorial". Nesse OND, seguindo a hierarquia de tópicos do documento, a primeira Estratégia de Defesa é o "Fortalecimento do Poder Nacional" e para tanto, as duas primeiras Ações Estratégicas de Defesa orientam o desenvolvimento do setor cibernético e a contribuição para o incremento do nível de segurança de estruturas críticas, em especial do sistema de água e de energia (BRASIL, 2016).

No setor de energia, destacam-se a Itaipu Binacional, em Foz do Iguaçu, responsável por fornecer cerca de 8,7% do consumo de energia do país[1] e o Centro Nuclear Almirante Álvaro Alberto, em Angra dos Reis. Tais estruturas, junto às demais usinas do Sistema Elétrico Brasileiro, constituem alvos de interesse cibernético cujos ataques bem-sucedidos são difíceis de prever, recuperar e estimar o impacto (NAFEES *et al.*, 2023).

2.2 Redes industriais e ICS

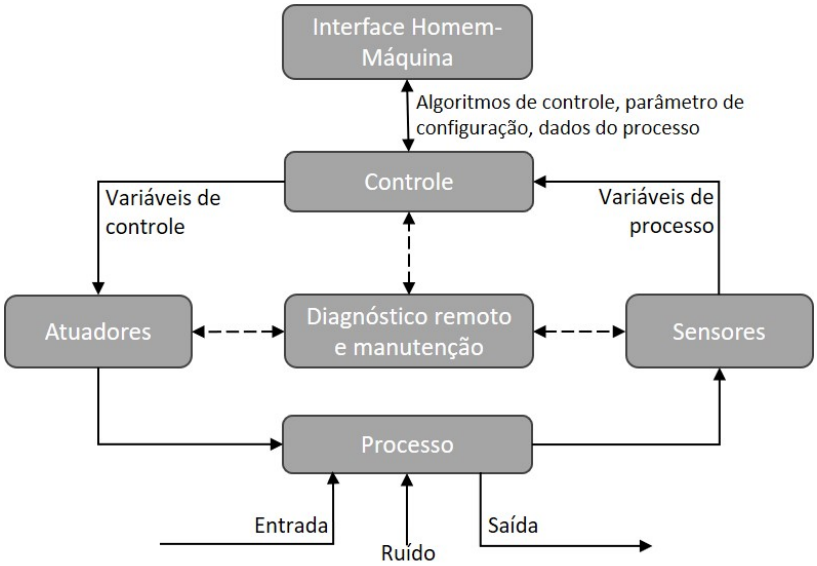
Além dos ambientes tradicionais de TI, muitas infraestruturas críticas abrangem ambientes TO (*Operational Technology*), que abarcam uma gama de dispositivos de sistemas ciberfísicos (CPS, do inglês *Cyber Physical System*) para o monitoramento e controle de instrumentos de campo, eventos e processos (STOUFFER, 2023). Exemplos típicos são Sistemas de Controle Industrial, sistemas de automação e robótica e sistemas de controle de acesso físico. A Figura 1 traz uma representação de uma rede TO típica, sem a sua conexão com o ambiente TI.

O termo "redes industriais" embute, apesar da aparente sutileza, diferenças expressivas quando comparado ao termo usual de "redes de computadores", presente em escritórios, universidades, repartições públicas, provedores de *internet*, *datacenters* e até mesmo nas redes domésticas. Enquanto essas redes contam com dispositivos bem conhecidos não apenas pelos profissionais técnicos mas também pelos usuário comum (e.g. estações de trabalho, celulares, roteadores, *switches*, impressoras, *firewalls*), nas redes industriais somam-se dispositivos do ambiente de campo, como sensores, atuadores, controladores lógicos programáveis, interfaces homem-máquina e sistemas *Supervisory Control and Data Acquisition* (SCADA). Já se fala, inclusive em *Industrial Internet of Things* (IIoT) para classificar sensores e máquinas com conectividade à *internet* com vistas a aprimorar processos industriais e fornecer dados em tempo real como parte de requisitos de negócio (STOUFFER, 2023).

Em um parque industrial classificado como infraestrutura crítica pode ser utilizada a arquitetura hierárquica *Purdue Enterprise Reference Architecture* (PERA), representada na Fi-

[1] https://www.itaipu.gov.br

Figura 1 – Fluxo típico de uma rede OT.



Fonte: Adaptado de Stouffer (2023).

gura 2, para modelar e visualizar as interações entre a rede corporativa (TI, níveis 4 e 5) e de campo (TO, níveis 0 a 3) (KOAY *et al.*, 2022). No nível zero, têm-se sensores e atuadores, instrumentos de medidas, válvulas e máquinas da ponta da linha de produção. No nível um estão os Controladores Lógicos Programáveis (PLC, do termo usual em inglês *Programmable Logic Controller*) e Unidades Terminais Remotas (RTU, de *Remote Terminal Unit*).

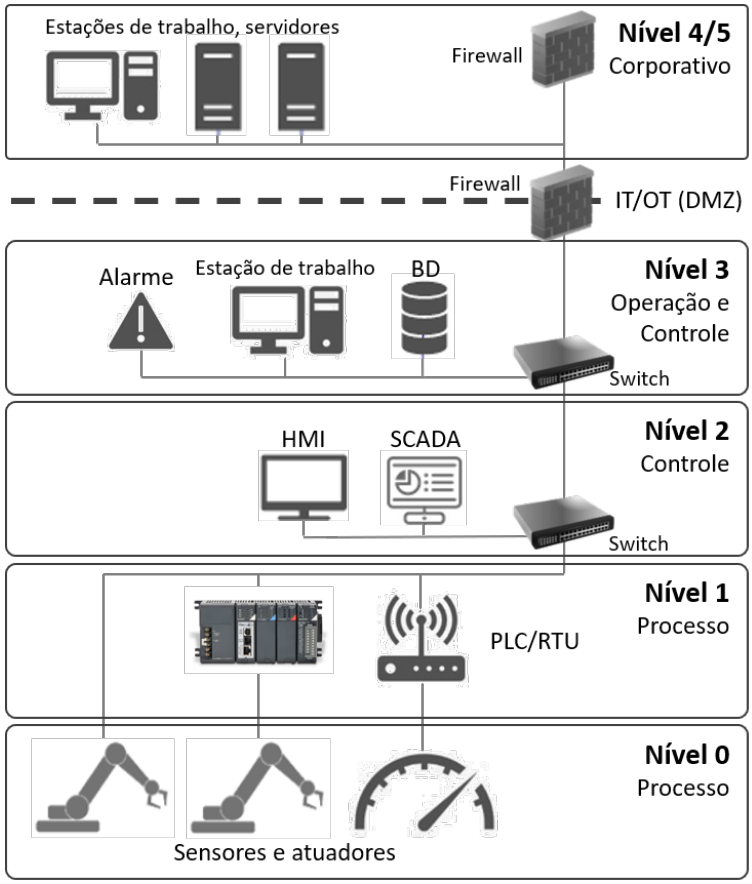
Os PLCs geralmente são programados em linguagem Ladder e continuamente coletam dados para tomadas de decisões e encaminhamento ao nível superior, operando em três etapas: a leitura e entrada dos dados dos equipamentos do nível zero, a execução de uma rotina armazenada em memória, e a escrita e emissão de comandos (ALANAZI; MAHMOOD; CHOWDHURY, 2023).

As RTUs possuem menor capacidade de processamento que os PLCs, não suportando *loops* de controle e algoritmos, sendo mais adequadas para comunicação sem fio e áreas geográficas esparsas, servindo de interface entre os dispositivos do nível zero e nível dois (YAACOUB *et al.*, 2020). Tal como os PLCs, são vulneráveis a ataques de desvio de autenticação, manipulação de dados e mensagens mal formadas (GRAHAM; HIEB; NABER, 2016).

No nível dois, as interfaces homem-máquina (HMI, de *Human-Machine Interface*) são dispositivos presentes na planta industrial que permitem, além da rápida visualização do estado da operação, em tempo real e por meio de interface gráfica específica para a aplicação, o ajuste de parâmetros de configuração dos processos monitorados (KOAY *et al.*, 2022). Podem funcionar por meio de um servidor web interno ou softwares proprietários da aplicação industrial (STOUFFER, 2023).

Os sistemas SCADA utilizam um computador central com o *software* que agrega as funções desse sistema, em quatro grupos: supervisão e controle, aquisição de dados, comunicação e armazenamento. O grupo de supervisão e controle é o responsável pela centralização

Figura 2 – Arquitetura de um ICS baseado no modelo PERA.



Fonte: Adaptado de Williams (1994).

das informações dispersas na linha de produção, integrando as HMIs com os demais grupos. O grupo de aquisição de dados contém os PLCs e RTUs e o grupo de armazenamento a interface com o banco de dados do nível três. O grupo de comunicação é composto pelos diversos protocolos empregados entre os demais grupos e seu componentes, bem como pelos meios físicos utilizados (ALANAZI; MAHMOOD; CHOWDHURY, 2023).

O banco de dados presente no nível três não contém dados relativos ao negócio ou à rede corporativa, mas sim séries temporais dos dados coletados pelos sensores, ações dos atuadores, decisões dos PLCs, enfim, variáveis de processo, de controle e parâmetros de configuração tratados pelo sistema SCADA. O termo usual em inglês é *historian*, podendo estar implementado em computadores com sistemas operacionais Windows, Unix ou Linux, ou soluções proprietárias em casos específicos (STOUFFER, 2023). Segundo Alanazi, Mahmood e Chowdhury (2023), a maioria dos sistemas SCADA utilizam SQL para armazenamento dos dados temporais.

O nível três, apesar de também cumprir funções de monitoramento e controle, abrange uma área maior que o nível dois e pode possuir HMIs, servidores de alarme, de gerência de processos, e estações de trabalho específicas de ICS (*Engineering Workstations*), que permitem a atualização de processos por várias plantas (KOAY *et al.*, 2022). Tais estações são propensas a ataques por serem relevadas em termos de segurança, rodarem processos com privilégios de

administrador, permitirem acesso remoto e não contarem com *softwares* antivírus (ANTROBUS *et al.*, 2016).

Os níveis quatro e cinco compreendem a rede corporativa típica de ambientes TI, com seus servidores de email, DNS e *web*, banco de dados, impressoras, estações de trabalho e conectividade a computadores pessoais e celulares de funcionários. Nestas camadas estão os serviços que suportam os negócios e a organização, bem como soluções de segurança usual, como *softwares* antivírus, *firewalls* e *backups* (KOAY *et al.*, 2022).

Outro fator que diferencia as redes industriais são os protocolos empregados entre os dispositivos de campo. Considerando a elevada gama de fabricantes, cada um com seus padrões e por vezes com soluções criadas sob demanda, diversos protocolos são empregados, sendo os mais comuns o Modbus, PROFINET, S7COMM e DNP3 (KOAY *et al.*, 2022). Se por um lado isso é uma vantagem quando dificulta a ação de cibercriminosos pelo esforço cumulativo de exploração de vulnerabilidades, por outro é uma desvantagem quando há também o acúmulo de vulnerabilidades conhecidas e não tratadas (YAACOUB *et al.*, 2020).

O quadro 1 traz uma visão geral das principais diferenças apresentadas. Neste quadro, as redes industriais foram definidas pelos ICS e suas diferenças com as redes convencionais, visto que essas diferenças ajudam a entender as vulnerabilidades e dificuldades de emprego de estratégias de proteção cibernética.

2.3 Segurança cibernética em redes industriais

Apesar das infraestruturas críticas contarem com redes corporativas orientadas à segurança, as características de um ambiente industrial não permitem empregar eficientemente as mesmas estratégias genéricas adotadas em redes de computadores (CHEMINOD; DURANTE; VALENZANO, 2013). Huda *et al.* (2017) ressalta que uma das principais causas das falhas de proteção contra ataques provocados por *malwares* em CPS recai no fato de que atualizações de sistemas operacionais e de *softwares* anti-vírus não são adequados para operar nesses sistemas de controle.

Enquanto atualizações em redes TI podem contar com testes em ambientes de homologação previamente à adoção nos ambientes de produção, nos sistemas industriais ter linhas de produção alternativas nem sempre é viável e atualizações necessitam de planejamento minucioso, demandando tempo e custo (HUDA *et al.*, 2017).

Na rede corporativa, a confidencialidade dos dados tem maior prioridade que a integridade, que por sua vez, sobrepõe a disponibilidade dos dados. Já nos sistemas industriais o requisito de disponibilidade é mais prioritário que o de integridade e confidencialidade (CHEMINOD; DURANTE; VALENZANO, 2013). Portanto, a adoção de medidas de segurança deve considerar os impactos no desempenho de um ICS, refletindo em decisões de *trade-off* entre a oportunidade e a segurança dos dados operados (YAACOUB *et al.*, 2020).

Quadro 1 – Diferenças entre ICS/TO e rede convencional/TI

	Rede convencional/TI	**ICS/TO**
	Características dos sistemas	
Número de usuários	muito alto	limitado
Expectativa de vida do sistema	anos	décadas
Indisponibilidade dos sistemas	muitas vezes tolerável	raramente tolerável
Tolerância a falhas	média/alta	baixa/nenhuma
Protocolos de comunicação	de uso geral/abertos (i.e. TCP/ IP, UDP)	de uso específico/proprietários /orientados a tempo real
Sistemas operacionais	genéricos (i.e. Windows, Unix)	tempo real/embarcados /projetados sob demanda
	Manutenção e atualização	
Patches e *upgrades* de *software*	muito frequentes	raros ou inexistentes
Versões legadas	raras	frequentemente em uso
Frequência de alterações no hardware	média/alta	baixa/muito baixa
	Aspectos de segurança cibernética	
Disponibilidade de informação	alta/muito alta	baixa/muito baixa
Auditorias	frequentes	raras/muito raras
Uso de antivírus	massivo	raro/inexistente (por vezes impossível)
Disponibilidade e adoção de *firewalls* e IDS	frequente/muito frequente	rara/por vezes impossível

Fonte: Adaptado de Cheminod, Durante e Valenzano (2013).

Conforme visto na seção 2.2, as diversas diferenças não apenas de equipamentos e protocolos como também de paradigmas e requisitos de negócio trazem dificuldades na adequação de técnicas de proteção cibernética, principalmente pelo fato de, historicamente, os sistemas industriais não serem projetados com atenção nesse tópico (CHEMINOD; DURANTE; VALENZANO, 2013). Uma das principais razões pela qual a segurança não é requisito quando do projeto de sistemas industriais é a percepção de que a segurança é um adendo que pode ser implementado posteriormente ou quando a necessidade surgir (SPAFFORD; METCALF; DYKSTRA, 2023).

Tendo em vista que tais sistemas compõem infraestruturas críticas de setores heterogêneos (e.g. gás e óleo, energia, água), progressivamente foram propostas variadas tentativas de padronização e de recomendações de segurança, voltadas para as necessidades específicas desses ambientes (DZUNG *et al.*, 2005).

Dentre os pontos em comum dessas tentativas, Cheminod, Durante e Valenzano (2013) afirmaram que as principais publicações de padronização convergem para a observação de

alguns conceitos conhecidos de Segurança da Informação, dentre os quais os termos “risco”, “vulnerabilidades” e “ameaça”. Em seu *survey*, classificaram os esforços de metodologia de segurança em três linhas de defesa em profundidade: prevenção, detecção e reação (CHEMINOD; DURANTE; VALENZANO, 2013).

A prevenção pode atuar na comunicação entre dispositivos buscando aprimorar a criptografia e processos de autenticação, enquanto a detecção pode analisar dados trafegados e aplicar técnicas de reconhecimento de padrões para alertar sobre comportamentos anômalos na rede (RADOGLOU-GRAMMATIKIS; SARIGIANNIDIS, 2019). A detecção pode valer-se de um IDS, já conhecido nas redes TI, e as reações ou são dependentes da detecção para então acionar alarmes e procedimentos de emergência (PAN; ADHIKARI, 2015) ou como medida de contenção de danos no esforço de conter o “efeito cascata”, em que as falhas vão se espalhando entre equipamentos e sistemas (CHEMINOD; DURANTE; VALENZANO, 2013).

Ainda que ataques cibernéticos sejam comuns na rede corporativa, essa não era ligada diretamente aos instrumentos de campo, cabendo ao ICS essa ponte. Portanto, tornam-se alvo de maior importância que os dispositivos da rede corporativa, em um contexto de exploração e ataque (CHEMINOD; DURANTE; VALENZANO, 2013). A heterogeneidade de fabricantes de dispositivos resulta na adição de vulnerabilidades e a intolerância a falhas e dificuldade de recuperação contribuem para o efeito cascata, com possíveis impactos catastróficos (YAACOUB *et al.*, 2020).

Dentre os esforços para prover soluções de segurança cibernética para redes industriais, foram observadas diferentes propostas na literatura, desde o reforço sistemático de práticas já consagradas em ambientes TI para a rede corporativa (BUCZAK; GUVEN, 2016), como a pesquisa em soluções orientadas especialmente para os ICS (YAACOUB *et al.*, 2020). Apesar da progressão de pesquisa e publicações nessa área, Mubarak *et al.* (2021) pontua que os dados da rede dos instrumentos de campo não são fáceis de obter, tanto pelo custo de construção dessas redes para pesquisa, quanto pela restrição de acesso aos dados pelas entidades responsáveis.

Os ataques orientados a manipulação dos dados operados pelos dispositivos de campo, geralmente do tipo de FDI, podem ser mitigados por reconhecimento de padrões aplicados em IDS específicos da rede industrial, com significativa vantagem sobre as mesmas técnicas empregadas apenas na rede corporativa (CHEMINOD; DURANTE; VALENZANO, 2013).

Outra vantagem da análise dos dados do ICS decorre do fato de que, ainda que a origem dos ataques esteja gradualmente aumentando de interna para externa, a ameaça de agentes internos (sabotagens por empregados insatisfeitos), com acesso aos instrumentos de campo, pode ignorar os mecanismos de defesa da rede corporativa mas seriam captados pelo IDS exclusivo do ICS (CHEMINOD; DURANTE; VALENZANO, 2013).

Os IDS podem ser classificados quanto às técnicas de detecção. Duas classes são frequentes na literatura: IDS baseados em assinatura e IDS baseado em anomalias. Nafees *et al.* (2023) apresentaram outras classes mais recentemente conceituadas, como a híbrida

entre assinatura e anomalia, no entanto as duas classes base são mais consolidadas, enquanto a híbrida continua sendo tema de pesquisa (NAFEES *et al.*, 2023).

A detecção baseada em assinatura emprega métodos de reconhecimento de padrões para identificar ataques previamente conhecidos, monitorando os dados e acionando o alarme quando uma sequência de dados é compatível com o espaço de assinaturas (NAFEES *et al.*, 2023). Resulta em baixos níveis de falsos alarmes, mas possui limitações no reconhecimento de ataques inéditos (conhecidos como *zero-day attacks*). Por definição, cria-se uma dependência de frequentes atualizações da base de dados de assinaturas, do contrário, em um ano, pode-se observar a queda de até 23% no desempenho de detecção (VIEGAS; SANTIN, 2020).

Na detecção baseada em anomalias, em vez do IDS aprender os padrões de assinaturas de ataques, o que é observado é o comportamento normal do sistema: qualquer desvio é considerado uma anomalia, provavelmente resultante de um ataque (ALANAZI; MAHMOOD; CHOWDHURY, 2023). Enquanto nas redes TI empregam-se IDS baseados em assinatura, diante da abundância de assinaturas disponíveis, nos ICS o apelo pelo uso de IDS baseado em anomalias é maior, a despeito da elevada taxa de falsos alarmes (NAFEES *et al.*, 2023).

Apesar da literatura concordar que há um progresso na questão da segurança, principalmente quanto à sensibilização e percepção da importância das infraestruturas críticas, a menção a uma equipe específica de segurança cibernética, como uma parte destacada do quadro de pessoal, ocorre somente nos artigos mais recentes. Nafees *et al.* (2023) verificaram que algumas terminologias são empregadas para designar essa equipe: *Security Operation Centre*, *Security Intelligence Centre* e *Cyber Security Operations Centre* (CSOC).

O papel de um analista do CSOC compreende analisar os incidentes de segurança e identificar ameaças aos ativos e serviços da organização a qual faz parte (SUNDARAMURTHY *et al.*, 2015). Em geral, o tratamento de incidentes é realizado segundo a prioridade identificada, de forma que alguns alertas podem levar mais tempo na fila para serem analisados do que outros de maior impacto e, logo, maior precedência (AGYEPONG *et al.*, 2020). Naturalmente, ferramentas de segurança auxiliam o trabalho dos analistas, como os já citados IDS, que inclusive podem ter assinaturas customizadas por analistas mais experientes (THOMAS, 2016).

2.4 Aprendizado de Máquinas

Nafees *et al.* (2023) em seu *survey* observaram o amplo emprego de Aprendizado de Máquinas e sua subárea *Deep Learning* nos IDS baseados em anomalia para emprego em redes industriais. Dentre os algoritmos tradicionais os trabalhos revisados citam *Support Vector Machine*, *Bayes* e seus derivados, *K-Nearest Neighbors*, *Decision Tree* e *Random Forest*. Já na área de *Deep Learning*, os recorrentes, para citar alguns, são as Redes Neurais Convolucionais, *Restricted Boltzman Machine*, *Long Short-Term Memory* e *Autoencoders*. Neste trabalho,

o termo “Aprendizado de Máquinas” é utilizado para tratar apenas dos métodos tradicionais, diferenciando daqueles do subgrupo de *Deep Learning*.

Uma abordagem típica de detecção baseada em Aprendizado de Máquinas geralmente consiste em duas fases, treino e teste, com a sucessão das seguintes etapas: identificação das classes de interesse, identificação do subconjunto necessário de atributos para a classificação, modelagem pelo aprendizado dos dados de treino, e uso do modelo para predição dos dados de teste (BUCZAK; GUVEN, 2016). As predições podem ser contestadas pelos valores verdade, quando disponíveis, permitindo a avaliação do desempenho do modelo segundo métricas diversas comentadas adiante.

Os dados para construção do conjunto de treino e teste variam segundo a aplicação, podendo ser extraídos dos pacotes de rede (e.g. *flags* de protocolos, endereços de origem e destino, dados de tráfego), do sistema operacional (e.g. chamadas de sistema, abertura de arquivos, cadeia de caracteres no *dump* de memória), dos dispositivos de campo (e.g. temperatura, pressão, estado, rotação), entre outros. O emprego de IDS baseados em Aprendizado de Máquinas é consolidado nas redes TI, com bons resultados e diversas publicações (BUCZAK; GUVEN, 2016). Já para as redes TO, suas peculiaridades impõem restrições e dificuldades, principalmente quanto à disponibilidade de dados, sendo as simulações a alternativa usual (ALANAZI; MAHMOOD; CHOWDHURY, 2023).

Como retrata o Quadro 1, os ICS tendem a ser mantidos em operação ininterruptamente e não toleram falhas ou pequenos distúrbios, de modo que experimentações e testes no ambiente de produção não são recomendados. Tanto para a comunidade acadêmica quanto para os órgãos responsáveis por infraestruturas críticas a construção de um laboratório para pesquisa em segurança cibernética de redes industriais é dificultada tanto pelo custo financeiro quanto pela falta de especialistas na área (MUBARAK *et al.*, 2021).

Foi observado, na Figura 2, a existência de um banco de dados para armazenamento dos registros da operação controlada pelos sistemas SCADA, contudo, quando esses dados existem eles não são facilmente liberados ao público pelos órgãos responsáveis pelo ambiente industrial, mesmo para pesquisa, devido à sensibilidade e criticidade das informações que podem ser extraídas (MUBARAK *et al.*, 2021). Dentre as possibilidades para contornar a falta de dados de ICS para experimentos uma das mais empregadas é a simulação virtual da rede industrial e de seus componentes. Trata-se de uma solução viável do ponto de vista dos custos, já amadurecida se avaliarmos o extenso emprego em redes TI, de acesso razoável, porém pouco precisa em relação aos CPS. Por outro lado, a recriação de um ambiente real para testes, apesar de mais versátil e fiel para os experimentos, é extremamente custosa (KUMAR; CHOI, 2022).

Alguns *datasets* são recorrentes na literatura, mesmo já desatualizados, citados com ressalvas e desenhados para redes TI, como o KDD99 e DARPA (MUBARAK *et al.*, 2021). Dentre os *datasets* disponíveis orientados a ICS uma das dificuldades é encontrar um que seja adequado para o ambiente de interesse, dado a miríade de dispositivos, aplicações e setores

que, quando analisados, não podem ser ignorados em prol da generalização (CHOI; YUN; KIM, 2018). Por exemplo, o *dataset* SWaT (*Secure Water Treatment*) contém dados dos dispositivos de campo trafegando mensagens com protocolo Modbus em simulação de uma estrutura de tratamento de água, o que pode dificultar o seu uso em aplicações com cenários que utilizem outros protocolos, mesmo se estes também forem de tratamento de água (GOH *et al.*, 2017).

Nesse sentido, Kumar e Choi (2022) afirmam como uma vantagem o uso de *Hardware-in-the-Loop* (HIL) para a construção de um *dataset* para ICS pela aproximação de um sistema real quando comparados com as simulações puramente digitais (KUMAR; CHOI, 2022). Para este trabalho, considerando a criticidade de infraestruturas críticas, em especial do setor de energia, e a proximidade geográfica com a usina Binacional Itaipu, foi escolhido o *dataset* HAI, a ser visto com mais detalhes no Capítulo 3.

De posse dos dados de origem, uma etapa de seleção de atributos (ou *features*, termo usual em inglês) visa extrair o subconjunto ótimo dentre os diversos atributos para não apenas eliminar dados irrelevantes, repetidos ou de pouca variância, mas também reduzir a dimensionalidade dos dados para obter ganhos de eficiência sobre o esforço computacional no tratamento de grandes volumes de dados (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006). Acrescenta-se o pré-processamento para preencher eventuais lacunas por dados faltantes ou corrompidos, normalizar ou escalonar os dados, remover ruídos ou aplicar técnicas de transformação dos dados como discretização (MAHDAVIFAR; GHORBANI, 2019).

A tarefa de classificação pode ser dividia em três grupos: supervisionada, semi-supervisionada e não supervisionada (BUCZAK; GUVEN, 2016). A tarefa de classificação supervisionada visa identificar corretamente um registro de dados como integrante de um ou mais classes previamente definidos, tendo por base o aprendizado em registros previamente rotulados e, em tese, representativos do problema avaliado. Podem haver duas ou mais classes e os registros podem ser rotulados em apenas uma ou mais classes. Nas aplicações de segurança o problema binário (duas classes mutuamente exclusivas) é o usualmente empregado, em que as classes são “ataque” e “normal” (KOAY *et al.*, 2022). O aprendizado supervisionado necessita da intervenção humana para rotular os registros nas classes de interesse e está intimamente relacionado aos IDS baseados em assinatura, não sendo eficientes para ataques inéditos ou que não constavam do conjunto de treino (PAN; ADHIKARI, 2015).

No aprendizado não supervisionado, trabalha-se com instâncias de dados não rotulados, sendo comum em tarefas de clusterização e redução de dimensionalidade (MUBARAK *et al.*, 2021). A classificação não supervisionada emprega o aprendizado de padrões e estruturas intrincadas dos dados de treino, não rotulados, para avaliar os dados de teste e tentar classificá-los segundo o modelo aprendido (HUDA *et al.*, 2017). Dessa forma, relacionada com os IDS baseados em anomalias, dispensa a intervenção humana no esforço de manualmente classificar as instâncias, além de se adequar a diversas situações reais, em que é comum a ausência de rótulos (KOAY *et al.*, 2022). Apesar da classificação não supervisionada não desempenhar tão

bem quanto a classificação supervisionada em termos de eficiência e detecção (ALMALAWI *et al.*, 2014), traz a vantagem de ser capaz de reconhecer ataques *zero-day* (SUDAR *et al.*, 2020).

O conceito de classificação semi-supervisionada situa-se entre a supervisionada e a não supervisionada, utilizando tanto registros rotulados quanto não rotulados (ENGELEN; HOOS, 2019). Para Mahdavifar e Ghorbani (2019), na detecção de anomalias semi-supervisionada o conjunto de treino contém dados rotulados apenas da classe de comportamento normal. Já segundo (BUCZAK; GUVEN, 2016), a coexistência no conjunto de treino de dados rotulados e não rotulados caracteriza o aprendizado semi-supervisionado, como um reforço na melhoria da classificação. Almalawi *et al.* (2014) assumem que os dados de treino pertencem apenas a uma classe, não necessariamente do comportamento normal, comentando ainda a desvantagem dessa abordagem residir na necessidade de longo tempo de operação em condições normais para a coleta de dados do comportamento normal, ficando sujeito a anomalias durante o processo, e na impossibilidade de coleta de dados de um sistema em constante estado de anomalia, sem garantia de cobrir os casos futuros (ALMALAWI *et al.*, 2014). Em seu *survey*, Engelen e Hoos (2019) resumem que a classificação semi-supervisionada consiste na flexibilização do uso de dados não rotulados quando a maioria é rotulado e vice-versa, objetivando ganhos de desempenho.

Apesar de classificação binária bem representar o problema da detecção de ataques no contexto de segurança cibernética, seja em ICS ou na rede TI, há outra abordagem que também visa a classificação mas que possui uma diferença sutil, denominada Detecção de Anomalias, Detecção de Novidades, *Outlier Detection* (OD), termo usual em inglês, ou ainda problemas de classificação de classe única (*One-class Classification*), a diferença de termos está relacionada às diferentes aplicações, sendo utilizados concomitantemente para designar a mesma ideia (TAX, 2001). A diferença reside no fato de não haver instâncias de ataques nos registros de treino, de modo que o modelo aprende apenas o comportamento normal e busca identificar as amostras divergentes do padrão aprendido (LAI *et al.*, 2021). Foi comum observar nos trabalhos revisados a correlação do uso desses termos com classificação não supervisionada, contudo, sem menções a diferença dos demais problemas de classificação. Pedregosa *et al.* (2011) ressaltam a diferença entre os termos detecção de novidade e detecção de *outliers*, indicando que naquele, semi-supervisionado, o foco está em classificar dados novos, externos ao conjunto de treino, enquanto nesse, não supervisionado, o foco está em diferenciar dados do mesmo grupo de observação (PEDREGOSA *et al.*, 2011). A maioria dos algoritmos de OD são considerados como não supervisionados tendo em vista o alto custo envolvido para rotular os dados (ZHAO *et al.*, 2019). Neste trabalho, o termo detecção de anomalias é utilizado para tratar tanto de OD quanto de detecção de novidades, por meio de algoritmos não supervisionados.

Identificada a natureza do problema a ser tratado e de posse dos dados tratados segue-se a escolha do modelo a ser empregado para aprendizado e predição. Não é trivial supor com antecedência qual modelo é mais adequado para determinado conjunto de dados. Apesar dos modelos se basearem em diferentes conceitos, estratégias e métodos, é possível, com poucas

adaptações, utilizar mais de um para o mesmo conjunto de dados, principalmente quando os modelos são do mesmo *framework* (KUMAR; CHOI, 2022). É comum observar nos experimentos o uso de diversos modelos para, após o resultado e avaliação, concluir pelo mais adequado (HUDA *et al.*, 2017). Koay *et al.* (2022) observaram que os melhores algoritmos para segurança cibenética de ICS são os baseados em árvores (*Random Forest* e *Decision Trees*), *Autoencoders*, *Deep Neural Networks* e *Deep Belief Networks*. Kumar e Choi (2022) verificaram que o desempenho dos algoritmos é dependente do conjunto de dados utilizados, mas que em termos gerais os modelos baseados em árvore são mais adequados no geral.

Nos trabalhos verificados que utilizam algoritmos de aprendizado de máquinas para a tarefa de classificação, quando mencionados, os hiper-parâmetros costumam ser configurados com os valores padrão de seus *frameworks*. Há uma variedade de fatores que influenciam na execução do algoritmo e resultado da detecção, como a seleção de *datasets*, partição dos conjuntos de treino e teste e o *tuning* (ENGELEN; HOOS, 2019). Idealmente a otimização dos parâmetros é orientada para os dados trabalhados, contudo o uso de valores fixos pré-definidos permite verificar a flexibilidade do modelo para adoção em outros *datasets* (TAX, 2001). Em geral, algoritmos otimizados são muito específicos para o conjunto de dados trabalhados, sendo pouco adequados para uso em outros conjunto de dados (ENGELEN; HOOS, 2019).

Após o aprendizado dos dados de treino pelos modelos escolhidos, os algoritmos realizam as predições dos dados de teste, classificando cada instância em ataque/anomalia ou em não ataque/comportamento normal. De posse das predições e dos rótulos reais do conjunto de treino, para um problema de classificação binário no contexto de detecção de anomalias causadas por ataques cibernéticos, há quatro possibilidades de resultados (SUDAR *et al.*, 2020):

- Verdadeiros Positivos (VP): registros de ataques corretamente detectados;
- Verdadeiros Negativos (VN): registros de comportamento normal não identificados como ataques;
- Falsos Positivos (FP): registros de comportamento normal incorretamente detectados como ataques; e
- Falsos Negativos (FN): registros de ataques não detectados, i.e. classificados como de comportamento normal.

Essas medidas permitem a construção de métricas para avaliação do desempenho do modelo treinado e aplicado, resultando em diferentes valores e oferecendo diferentes semânticas, sendo esse último fator de grande importância na escolha, dado que o uso de métricas inadequadas pode levar a decisões equivocadas na gestão de segurança da infraestrutura crítica (JUBA; LE, 2019). As quatro métricas mais usuais nos trabalhos de classificação binários via Aprendizado de Máquinas em aplicações de segurança para ICS são (KOAY *et al.*, 2022):

- Acurácia: $(VP + VN)/(VP + VN + FP + FN)$. É a mais simples e provavelmente a mais empregada, útil para problemas de classes balanceadas mas frequentemente empregada também para problemas de classes não balanceadas. Almeja representar a medida da correta classificação de ambas classes, ataque e comportamento normal. Como não considera a proporção entre as classes, pode levar a entendimentos incorretos do desempenho do modelo, quanto mais desbalanceado for o problema. Os problemas reais, em geral, são de classes desbalanceadas (HOI *et al.*, 2021). Para ilustrar, se o conjunto de teste possuísse apenas 1% de registros de ataques e o classificador detectasse que todos os registros são de comportamento normal, i.e., que não houve ataques, o valor da métrica acurácia seria de 99%, passando a impressão de que o modelo apresentou alto desempenho, apesar de falhar em detectar o ataque. Ainda, se para o mesmo conjunto de dados outro classificador encontrasse dois falsos positivos além do registro de ataque, resultaria em acurácia de 98%, menor que o modelo anterior. Apesar desse conhecido problema, a métrica Acurácia ainda é amplamente empregada, seja pela facilidade de entendimento ou de cálculo (KOAY *et al.*, 2022).

- Precisão: $VP/(VP + FP)$. Traduz a capacidade da modelagem em não rotular como anormalidade um registro de comportamento normal. Quanto maior seu valor, fixado VP, menos um modelo produz falsos positivos. Também é dito como a porcentagem de amostras de ataques corretamente classificadas (AL-ABASSI *et al.*, 2020). No entanto, apesar de traduzir a capacidade de não gerar falsos positivos, por sua construção, um modelo pode apresentar elevada taxa de precisão mesmo com considerável número de ataques não detectados (ALMALAWI *et al.*, 2014).

- *Recall* ou Sensitividade ou Taxa de detecção: $VP/(VP + FN)$. Reflete a habilidade do modelo em encontrar todas as anomalias. Quanto maior o seu valor, fixado VP, menos o modelo produz falsos negativos. Baixos valores de *recall* indicam que muitos ataques não foram detectados, o que é crítico em aplicações de segurança de infraestrutura crítica (RADOGLOU-GRAMMATIKIS; SARIGIANNIDIS, 2019). A escolha da métrica *recall* reflete uma opção conservadora pensando no cenário de pior caso (danos elevados resultantes de ataques bem-sucedidos), mas essa opção pode levar a um elevado número de falsos-positivos, que demandam análise posterior (CHEMINOD; DURANTE; VALENZANO, 2013).

- *F_1-score*: $2VP/(2VP + FP + FN)$. Média harmônica entre a Precisão e o *Recall*, junto da Acurácia é uma das métricas mais empregadas em problemas de classificação binária (KOAY *et al.*, 2022). É recomendada para tratar de problemas de classes desbalanceadas quando comparadas às demais métricas (JENI; COHN; TORRE, 2013).

Além das quatro métricas mais usuais, diversas outras, de fórmulas mais simples ou complexas, foram propostas para avaliar classificadores em aplicações de detecção de anomalias (ALMALAWI *et al.*, 2014). Neste trabalho, é utilizada também a AUC, representada pela Equação 1, que expressa a intuição da probabilidade de um registro aleatório de ataque ter um escore maior que um registro de comportamento normal, escore esse fornecido pela função de decisão do classificador empregado, quando da predição do conjunto teste (CALDERS; JAROSZEWICZ, 2007). Seja $C(r)$ a classe de um registro r, sendo, para um problema binário, $C(r) = 0$ para indicar a classe de comportamento normal e $C(r) = 1$ para indicar a classe de um ataque, a métrica AUC de um classificador f pode ser definida como

$$AUC(f) := P(f(x) < f(y)|C(x) = 0,C(y) = 1) \tag{1}$$

A contagem de FN e FP dão origem a outras duas métricas: a taxa de falsos negativos (FNR) e taxa de falsos positivos (FPR), formuladas na equações 2 e 3, respectivamente. Koay *et al.* (2022) constataram em seu *survey* que apenas quatro de trinta artigos anotaram a contagem de falsos positivos e falsos negativos, no entanto esses valores podem ajudar tanto na comparação entre métricas quanto entre algoritmos. Entre cenários distintos, a opção pelas métricas FNR e FPR permitem uma comparação normalizada uma vez que trazem a informação percentual para cada cenário.

$$FNR = \frac{FN}{VP + FN} \tag{2}$$

$$FPR = \frac{FP}{VN + FP} \tag{3}$$

3 TRABALHOS RELACIONADOS

Neste Capítulo, são apresentados os trabalhos que deram origem à proposta. No Capítulo 2, foi visto que as infraestruturas críticas, essenciais para o bem-estar da sociedade e alvo de cibercriminosos, especialmente do setor de energia, são compostas por redes industriais TO cada vez mais interconectadas com a rede corporativa TI. Essas redes industriais possuem peculiaridades que as diferenciam das redes TI, trazendo dificuldades para implementação de estratégias de segurança. Foi observado que o uso de IDS baseados em Aprendizado de Máquinas, quando monitorando os dados do ICS, complementam as ações de segurança e apresentam resultados promissores.

3.1 HIL-based Augmented ICS Dataset

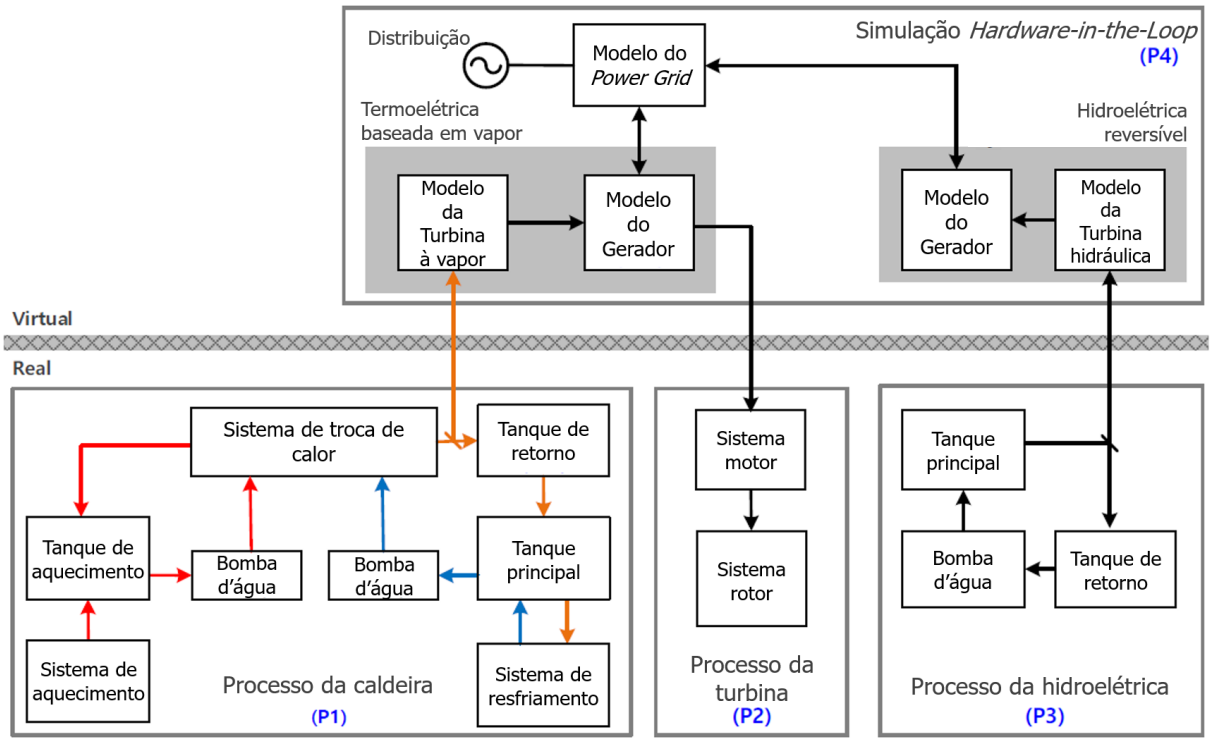
Para o contexto do Brasil, considerando a Binacional Itaipu, seria valoroso obter informações da planta e seus equipamentos, para a montagem de plataforma de simulação, ou ainda os dados da rede TO para aplicação em métodos de Aprendizado de Máquinas. Naturalmente esses dados e informações são restritos, contudo, foi firmada uma parceria entre a Itaipu, o Exército Brasileiro e a Fundação Parque Tecnológico Itaipu para consolidação do Centro de Estudos Avançados em Proteção de Estruturas Estratégicas, do qual fazem parte o Laboratório de Simulação Cibernética (LASC) e o Laboratório de Segurança Eletrônica, de Comunicações e Cibernética (LASEC²) (ITAIPU; EXÉRCITO; PTI-BR, 2022).

No Capítulo 2, foi ressaltada a dificuldade em construir *datasets* para ICS tanto para o comportamento normal quanto para a análise de ataques, sendo a simulação a alternativa mais adotada. Shin *et al.* (2020) propuseram uma plataforma para simular e emular uma usina de geração de energia elétrica mista de termoelétrica baseada em vapor e hidroelétrica reversível. Foram empregados componentes reais e não apenas a simulação virtual, um grande diferencial em comparação aos *datasets* disponíveis (KUMAR; CHOI, 2022). A integração entre os componentes reais e a simulação virtual é feita por meio de *Hardware-in-the-Loop*, de onde vem o nome da plataforma e de seus *datasets*: HAI (SHIN *et al.*, 2020).

A proposta do HAI vem em resposta a um dos principais desafios na pesquisa de segurança cibernética em ICS, conforme ressalta Nafees *et al.* (2023): o uso de HIL em simulações complexas de um ambiente industrial para compreender os efeitos de ataques, em especial nos sistemas de comunicação e de geração de energia.

A plataforma é dividida em quatro processos, ilustrados na Figura 3. O processo da caldeira (P1) emprega um *boiler* Emerson Ovation responsável pela transferência de calor e quatro controladores para o nível, temperatura, vazão e pressão da água no sistema. Os valores de temperatura e pressão do vapor dão entrada no simulador do modelo da turbina, que faz a interface com o processo da turbina (P2). O processo P2 é composto por um *kit* de turbina GE's Mark VIe e é capaz de monitorar e controlar a velocidade de rotação e vibração do eixo rotor,

Figura 3 – Diagrama da plataforma HAI.



Fonte: Adaptado de Shin *et al.* (2023).

sincronizado com o modelo simulado no HIL. O processo da hidroelétrica (P3) utiliza um PLC Siemens S7-300 integrado com o controle do nível da água e bomba d'água do reservatório. Os dados de vazão, pressão e nível são encaminhados ao HIL que simula a geração e consumo de energia. Por fim, o processo P4 é o responsável pela integração dos demais por meio de dispositivos Siemens S7-1500 PLC e ET200, sendo constituído pelo HIL dSPACE SCALEXIO e permitindo a coleta e ajuste dos dados (SHIN *et al.*, 2023).

O HAI já possui quatro versões: 20.07, 21.03, 22.04 e 23.05, cada versão referente ao ano e mês da coleta dos dados. A Tabela 1 apresenta as informações das versões para rápida consulta. As capturas ocorrem a cada segundo e de forma sequencial, permitindo o uso dos dados tanto com amostragem quanto com séries temporais.

Os pontos de captura correspondem às *features* dos *datasets* e são classificados em variáveis de controle, de configuração ou de processo. As variáveis de configuração representam os parâmetros de execução e ajustes que seriam feitos pelo operador da HMI, como por exemplo limites de operação. Variáveis de processo são leituras de sensores ou sinais emitidos por atuadores em resposta a comandos ou rotinas programadas. Como exemplo citamos a leitura de temperatura do tanque de aquecimento do processo P1. Variáveis de controle são geradas pela modelagem da operação tendo por entrada as variáveis de processo e de configuração ou ainda de outras variáveis de controle. O sinal que opera a velocidade do motor da turbina para sincronia com o HIL é um exemplo (SHIN *et al.*, 2023).

Tabela 1 – Dados das versões do HAI

Versão	*Features*	Cenário Normal		Cenário de Ataque		
		Arquivo (.csv)	Duração (horas)	Arquivo (.csv)	Ataques	Duração (horas)
HAI 20.07	59	train1	86	test1	28	81
		train2	91	test2	10	42
HAI 21.03	78	train1	60	test1	5	12
		train2	63	test2	20	33
		train3	229	test3	8	30
				test4	5	11
				test5	12	26
HAI 22.04	86	train1	26	test1	7	24
		train2	56	test2	17	23
		train3	35	test3	10	17.3
		train4	24	test4	24	36
		train5	66			
		train6	72			
HAI 23.05	86	train1	78	test1	14	15
		train2	81	test2	38	64
		train3	35			
		train4	55			

Fonte: Extraído de Shin *et al.* (2023).

3.2 Cenários de ataques

O HAI foi proposto para preencher a lacuna de *datasets* orientados a segurança cibernética para CPS e portanto foi desenvolvido com foco nos efeitos de atividade maliciosa na rede TO (SHIN *et al.*, 2020). Nesse sentido foi desenvolvido com a premissa de que os ataques são furtivos, ou seja, alterações nos dados operados por sensores e atuadores não são triviais de identificar e não são representados por valores fixos, como observados em outros *datasets*. Outro diferencial diz respeito aos efeitos sistêmicos do ataque, i.e. ataques em variáveis de um dos processos pode gerar distúrbios nos demais, o que é mais próximo de sistemas reais. Por fim, foi pensado em automatizar a criação de cenários de ataques, como forma de aumentar a confiança nos rótulos dos registros de ataque, dispensando a atuação humana para rotular cada registro da plataforma (SHIN *et al.*, 2020).

Para a operação normal da plataforma foram desenvolvidas ferramentas automatizadas para simular o comportamento humano, em específico, dos operadores do ambiente industrial ao imputar parâmetros de execução nas HMI em atenção às informações apresentadas. Essas ferramentas operam em ambiente separado da plataforma, interagindo apenas no papel do operador segundo critérios pré-estabelecidos e com alguma margem de aleatorização, objetivando aproximar de um cenário real. Outra vantagem do uso da ferramenta é facilitar a replicação de experimentos e reduzir custos com a intervenção humana quando da operação continuada por longo tempo (SHIN *et al.*, 2020).

Segundo Liang *et al.* (2017), ataques bem-sucedidos do tipo FDI, em usinas e estações de energia, comprometem as leituras dos sensores de tal modo que erros se propagam e afetam os mecanismos de estimativa das variáveis, passando despercebidos pelos controles de segurança. Essas pequenas alterações podem induzir os operadores da HMI a imputar configurações equivocadas na linha de produção, resultando em desgastes prematuros, perda de eficiência, prejuízos econômicos, e, no pior caso, falhas críticas pelo acúmulo de falhas residuais (LIANG *et al.*, 2017).

Em termos de objetivo geral, esses ataques visam comprometer a integridade dos dados, que, diferente dos pacotes trafegados nas redes TI, compõem-se de leituras numéricas de componentes em geral do processo de estimativa de estado da usina de geração elétrica (MO; SINOPOLI, 2010). Quanto a objetivos específicos, Sayghe *et al.* (2020) enumeram os ataques de FDI quanto ao impacto:

- Impacto na operação da malha elétrica: resultado, por exemplo, de ataques de redistribuição de carga, ocupando recursos da estação destinados à redução de custo operacional com situações advindas de estados de operação corrompidos;
- Impacto no processo de roteamento da distribuição de energia: ocorre pela injeção de dados falsos na comunicação entre os nós fornecedores de energia, afetando a tomada de decisão para o balanceamento de cargas e resposta de demandas de potência e resultando em severo aumento dos custos de operação; e
- Impacto econômico: a flutuação de preços em tempo real segundo a demanda de carga e capacidade de produção pode ser afetada por FDI no processo de estimativa de estados e assim resultar em faturamentos incorretos e prejuízos à estrutura atacada.

Quadro 2 – Exemplo de cenário de ataque no *dataset* HAI

Cenário	Alvo			Descrição	Versão
	Controlador	Variável	Ponto de captura		
AP16	P1-LC	CV1	P1_LCV01D	Decremento ou incremento de variável de controle do controlador P1-LC. Volta à normalidade.	HAI 20.07 HAI 21.03 HAI 22.04 HAI 23.05

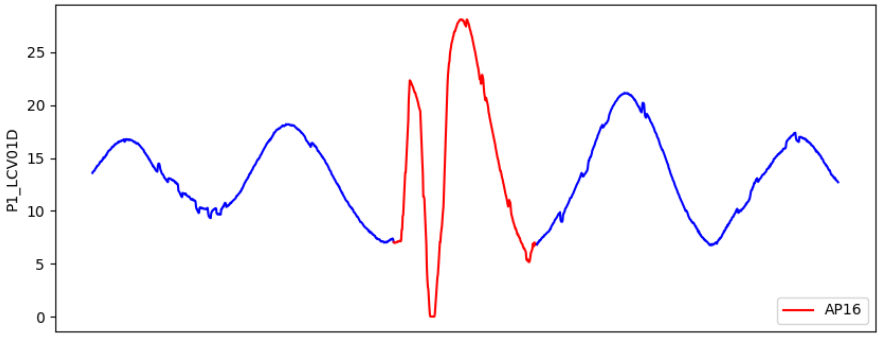
Fonte: Adaptado de Shin *et al.* (2023).

No HAI, os cenários de ataque podem afetar variáveis de controle, de processo ou de configuração, estes separados em parâmetros de controle e *setpoints*. Por construção, qualquer das *features* podem ser manipuladas, em tempo real na operação da plataforma, para simular um cenário de ataque do tipo FDI. Esses dados são tramitados pelo ICS e, quando manipulados, podem ou passar despercebidos ou causar distúrbios locais ou nos demais processos. Dessa forma, é obtida a característica da furtividade e do efeito sistêmico pretendidos para o *dataset*

(SHIN *et al.*, 2023). Apenas para facilitar a compreensão de um cenário de ataque, citamos um dos 37 descritos na última versão do HAI, ilustrado no Quadro 2.

No Quadro 2, P1-LC é o controlador do nível dos tanques de P1, CV1 refere-se a um grupo de variáveis de controle e P1_LCV01D é o ponto de captura referente ao comando de posição da válvula de acesso ao tanque de retorno da caldeira. A Figura 4 permite visualizar a leitura da *feature* alvo do ataque exemplificado, em um extrato da sequência temporal.

Figura 4 – Extrato de leitura da *feature* alvo de exemplo de ataque.



Fonte: Extraído de Shin *et al.* (2023).

Os cenários de ataque do HAI não são focados em detalhar como poderiam ser realizados, mas sim na efetiva corrupção dos dados trafegados, típica de FDI. São ataques sutis com efeitos progressivos e representam bem o desafio alertado por Cheminod, Durante e Valenzano (2013): mudança lenta e gradual do comportamento do sistema para maximizar os danos quando alcançar um estado de vulnerabilidade extremo, e pequenas perturbações persistentes por longo tempo. Mais de um ataque podem ocorrer simultaneamente, desde que em alvos distintos, e a proporção de registros de ataques em relação aos de comportamento normal, considerando arquivos de treino e de teste, é de 18.303 para 977.097 no HAI 20.07, 8.947 para 1.314.661 no HAI 21.03, 12.030 para 1.353.572 no HAI 22.04 e 11.384 para 1.169.416 no HAI 23.05. Em termos percentuais, na média, aproximadamente 1,0899% das instâncias registradas são da classe ataque.

3.3 Abordagens de detecção no HAI

Apesar da disposição dos arquivos do *dataset* direcionarem para o emprego em detecção de anomalias (arquivos de treino sem registros de ataques), os arquivos de teste são rotulados e permitem a adoção da detecção por assinatura. Tal possibilidade traduz uma vantagem e contribui na linha de pesquisa de detecção por assinatura em ICS, mesmo que o foco majoritário desses trabalhos recaem em detecção por anomalia com *Deep Learning* (ALANAZI; MAHMOOD; CHOWDHURY, 2023).

Dos *surveys* de segurança cibernética para ICS verificados, apenas o de Koay *et al.* (2022) apresenta o HAI como *dataset* de ICS para emprego em pesquisa de Aprendizado de

Máquinas, classificando em relação ao tipo de ataque cibernético como de FDI e colocando o trabalho de Mokhtari *et al.* (2021) como o de maior acurácia verificado para o *dataset*.

A proposta de Xingchao (2020) fundamenta-se em métodos de *Deep Learning* para detecção de anomalias em dados temporais com aprendizado não supervisionado. Denominaram seu *framework Stacked Gated Recurrent Unit - Infrequent Residual Analysis*. Afirmam que o modelo pode ser treinado em fluxo de dados brutos oriundos dos sensores e sem nenhum rótulo pré-estabelecido, uma das questões do problema tratado pela pesquisa. Empregaram o HAI-20 e uma versão derivada, disponível à época exclusivamente em uma plataforma sul-coreana para competições de Aprendizado de Máquinas[1]. No pré-processamento, os dados são normalizados e suavizados pela média móvel exponencial para redução de ruídos. O modelo proposto foi comparado pela métrica F_1*-score* com os resultados obtidos pelo algoritmo *Long Short-Term Memory*, obtendo valores até 5,9% superiores ao *baseline*, alcançando de 0.811 a 0.977 na métrica em termos absolutos (XINGCHAO, 2020).

No trabalho de Mokhtari *et al.* (2021), os autores advogam pelo uso dos dados trafegados na rede TO em complemento às medidas de proteção da rede TI, denominando a abordagem proposta como MIDS, acrônimo do inglês para sistemas de detecção de intrusão de medidas, em atenção aos dados medidos e tratados pelos sistemas SCADA. Utilizaram o HAI-20 para o aprendizado dos classificadores *K-Nearest Neighbors* (KNN), *Random Forest* e *Decision Tree*. Na etapa de seleção de *features* reduziram das 59 iniciais para 17 finais por meio da Correlação de Pearson. Os dados foram normalizados e, optando por lidar com o desbalanceamento das classes, empregaram o método *Synthetic Minority Over-sampling Technique* (SMOTE), o qual cria instâncias da classe desejada com base nos registros existentes (CHAWLA *et al.*, 2002). A proporção entre treino e teste foi estabelecida em 0,7 e 0,3, respectivamente, e a métrica escolhida para avaliação das predições foi a acurácia. O algoritmo *Random Forest* obteve o melhor desempenho, com métrica em 0,9976 (MOKHTARI *et al.*, 2021).

Pang, Pu e Li (2022) propuseram um algoritmo híbrido entre *Vector Quantization* (VQ) e *One-Class Support Vector Machine* (OCSVM) para problemas de detecção de anomalias. A abordagem almeja contornar a tarefa de seleção de *kernel* para o OCSVM, utilizando o VQ para mapear diretamente os dados em um hiper-espaço de *features*. Compararam o algoritmo proposto com outros seis algoritmos de OD utilizando as métricas *recall*, precisão e F_1*-score* em dois *datasets* (KDD-99 e HAI-20). Para os testes no HAI-20, utilizaram o processo P2 e repetiram a execução dez vezes, utilizando as médias dos resultados para concluir sobre o desempenho da proposta. Foram amostrados aleatoriamente 5.000 registros de treino e 1.500 de teste, dos quais 500 são de registros marcados como ataque. Os dados foram normalizados antes do aprendizado no intervalo $[-1,1]$. Obtiveram valores de 0,945 em *recall*, 0,996 em precisão e 0,97 para F_1*-score*, 2% superior ao segundo colocado (PANG; PU; LI, 2022).

Em Wang *et al.* (2023), os dados do processo P1 do HAI 21.03 foram utilizados como base para experimentos numéricos do *framework* proposto, objetivando a detecção de anoma-

[1] https://dacon.io/

lias com base em redes neurais de grafos espaço-temporais. O *framework* foi elaborado em atenção à dificuldade de se extrair, em redes industriais, *features* espaço-temporalmente correlacionadas com os métodos baseados em redes convolucionais de grafos. Os autores concluíram pela efetividade da proposta para o uso de monitoramento de ICS e modelagem dinâmica de grafos, redução de dimensionalidade e filtragem de ruídos. Na comparação com outras seis técnicas baseadas em grafos e redes neurais obtiveram resultados até 25,5% superiores na métrica F_1*-score* em relação ao segundo colocado. Em vez de utilizarem os cenários de ataques do *dataset*, simularam dez classes de falhas pelo aumento ou diminuição do tamanho dos valores registrados, justificando que tal comportamento é semelhante ao causado por desgastes de equipamentos e ataques na rede (WANG *et al.*, 2023).

Shin *et al.* (2023) relacionaram alguns dos trabalhos que fazem uso ou menção a plataforma, dentre os quais os recém-referenciados. O Quadro 3 traz referência dos demais que utilizaram os dados do HAI para detecção de anomalias nas áreas de Aprendizado de Máquinas, *Deep Learning* ou teoria dos Grafos.

Quadro 3 – Trabalhos relacionados ao HAI

Autor(es)	Estratégia	Abordagem
Xingchao (2020)	*Deep Learning*	Anomalias
Kim e Lee (2021)	*Deep Learning*	Anomalias
Mokhtari *et al.* (2021)	Aprendizado de Máquinas	Assinatura
Kim e Kim (2021)	*Deep Learning*	Anomalias
Kim *et al.* (2021)	*Deep Learning*	Anomalias
Kabore *et al.* (2021)	*Deep Learning*	Assinatura e anomalias
Bae, Hwang e Lee (2021)	*Deep Learning*	Anomalias
Demertzis *et al.* (2022)	*Deep Learning*	Anomalias
Seong *et al.* (2022)	*Deep Learning*	Anomalias
Xue e Yan (2022)	*Deep Learning*	Anomalias
Fu e Xue (2022)	*Deep Learning*	Anomalias
Han e Woo (2022)	Grafos	Anomalias
Pang, Pu e Li (2022)	Aprendizado de Máquinas	Anomalias
Wang *et al.* (2023)	Grafos	Assinatura

Fonte: Extraído de Shin *et al.* (2023).

Percebe-se uma lacuna nos trabalhos do HAI dado que a preferência parece ser pela abordagem *Deep Learning* e séries temporais. O próximo capítulo apresenta a proposta para o uso de algoritmos tradicionais de Aprendizado de Máquinas com aplicação no HAI em classificação supervisionada e não supervisionada, no intuito de verificar a viabilidade do emprego dessas técnicas em *dataset* de ICS simulado em HIL, podendo servir de base para futuras pes-

quisas pelos LASC e LASEC², visando aumentar os níveis de segurança da infraestrutura crítica Itaipu ou outras redes TO que possam ser simuladas nesses ambientes.

4 PROPOSTA

Foi verificado nos capítulos anteriores que as Infraestruturas Críticas são alvos de grande interesse para a exploração cibernética com diversas motivações. Tais estruturas comportam redes industriais baseadas em TO, com várias diferenças quando comparadas às redes baseadas em TI, que afetam aspectos de segurança cibernética. Dados como temperatura, pressão e tensão podem ser afetados por ameaças e resultar em desgastes prematuros de componentes eletromecânicos, efeito cascata e interrupção do serviço crítico. Uma das formas de detectar ataques que alterem esses dados compreende o uso de técnicas de Aprendizado de Máquinas, em especial a tarefa de classificação. Este capítulo apresenta a proposta para a detecção dos registros correspondentes a ataques nos dados simulados pelo HAI, a metodologia para implementar a proposta, as ferramentas e o modo de avaliação dos resultados. Os códigos-fonte empregados, hiper-parâmetros das funções e detalhes da implementação estão disponibilizados no *github*[1] para permitir a reprodução dos experimentos.

4.1 Arquitetura

Para a detecção de anomalias em ICS são utilizados algoritmos de Aprendizado de Máquinas para a tarefa de classificação binária em *dataset* representativo de uma rede industrial. Nos trabalhos relacionados ao HAI, foi observada uma inclinação para o emprego de técnicas contidas na área de *Deep Learning*, sugerindo espaço para pesquisa com os algoritmos tradicionais de classificação supervisionada (detecção por assinatura) e não supervisionada (detecção por novidade).

Em vez de optar por uma abordagem ou outra, de forma isolada, propõe-se a composição de classificadores supervisionados e não supervisionados, de modo a verificar ganhos nas métricas de detecção na comparação com a abordagem supervisionada.

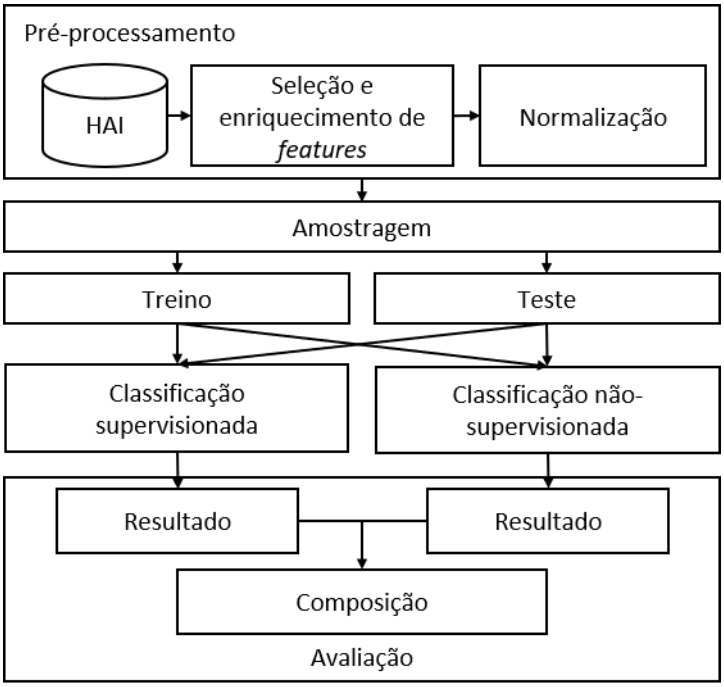
Neste trabalho, a existência de um CSOC em um órgão responsável por uma infraestrutura crítica é uma premissa que reflete na operação do serviço crítico. Considerando os custos envolvidos na contratação de especialistas e na resolução de incidentes cibernéticos, busca-se medidas que permitam desonerar essa equipe com a menor redução nos níveis de segurança.

A detecção em redes industriais como a representada pelo HAI complementam os mecanismos de segurança já implementados nas redes corporativas TI. Os dados registrados nas capturas do HAI traduzem o observado no ICS entre os sensores, PLCs, SCADA e HMI. Ataques sutis ou persistentes, sejam para o desgaste prematuro de partes eletromecânicas ou acúmulo de falhas culminando em falha crítica, podem passar despercebidas pelos IDS da rede corporativa mas detectados quando avaliados os dados de temperatura, pressão e demais da rede industrial.

[1] https://github.com/ivo-abreu/mestrado

Em termos gerais, foi proposto um sistema que, com base nos dados provenientes de uma rede TO de um ICS, por meio de técnicas de reconhecimento de padrões possam classificar corretamente os registros como ataques ou comportamento normal. O sistema proposto foi publicado em (NICOLAIO; MUNARETTO; FONSECA, 2023) e está apresentado na Figura 5.

Figura 5 – Sistema proposto.



Fonte: Nicolaio, Munaretto e Fonseca (2023).

A proposta surge da pretensão de detectar alterações nos dados numéricos trafegados na rede TO causadas como efeitos de ataques cibernéticos, sem a necessidade de especificar como ocorrem. Na primeira fase recorre-se às assinaturas conhecidas dos ataques e, portanto, emprega-se a abordagem supervisionada de Aprendizado de Máquinas. O classificador é treinado para reconhecer padrões já rotulados ou como ataque ou como de comportamento normal e realiza as predições nos dados de teste.

A segunda fase objetiva precaver-se de ameaças, que podem ser não apenas resultantes de ataques, como também de desgaste natural dos componentes, de acidentes operacionais ou de falhas humanas. Nesse caso, emprega-se a detecção de anomalias por meio do aprendizado não supervisionado na intenção de separar os dados de comportamento normal dos dados anômalos.

Considerando a criticidade dos impactos resultantes, quando no pior caso, de ataques em Infraestruturas Críticas, entende-se que é melhor errar por excesso que pela falta de cautela e, portanto, uma abordagem conservadora na composição das fases é a soma lógica das predições. Nesse caso, tem-se que uma instância é considerada um ataque caso seja detectada por qualquer uma das fases.

Da Figura 5 depreende-se quatro etapas: pré-processamento do *dataset*, amostragem e construção dos conjuntos de treino e teste, aprendizado dos modelos e predições, e composição

e avaliação dos resultados. Essas etapas serão descritas na metodologia para implementação da proposta.

4.2 Metodologia

4.2.1 Ferramentas

Para a implementação da proposta e execução dos algoritmos de Aprendizado de Máquinas, foi selecionada a biblioteca Scikit-learn na versão 1.2.0 (PEDREGOSA *et al.*, 2011). Esse *framework* é bastante conhecido e aceito na comunidade acadêmica, versátil e documentado, atendendo as necessidades da pesquisa. A linguagem de programação é Python na versão 3.8 e o ambiente de execução uma máquina virtual nas configurações do Quadro 4.

Quadro 4 – Descrição do ambiente de execução dos experimentos

Ferramenta/Configuração	**Descrição**
Processador	Intel 16 cores
Memória RAM	32 GB
Espaço em disco	100 GB
Sistema Operacional	Debian 10
Linguagem de programação	Python versão 3.8
Acelerador	Intelex versão 2023.0.1

Fonte: Autoria própria (2024).

O *dataset* escolhido para representar o ICS é o HAI, cuja descrição geral consta do Capítulo 3.

4.2.2 Pré-processamento

Na etapa de pré-processamento o ponto de partida é a escolha dos dados a serem trabalhados. Foi visto que o HAI consiste de quatro versões e que cada versão possui diferenças quanto aos pontos de captura e tempo total de captura. Também cada versão possui arquivos nominados `train#.csv`, gerados com a finalidade de simular capturas de comportamento normal, e arquivos nominados `test#.csv`, possuindo registros de atividade normal e cenários de ataque.

Tendo em vista a gama de dados disponíveis para o experimento, ainda mais se consideradas as adições e permutas de arquivos treino e teste (desde que da mesma versão), foram escolhidas aleatoriamente combinações de arquivos treino e teste de uma mesma versão, ou somente arquivos de teste. Dentre as versões do HAI e por uma questão de tempo disponível

para os experimentos, foram selecionadas as de 2020[2] e 2022[3]. A primeira pela presença majoritária nos trabalhos relacionados e a outra por ser, no início dos experimentos, a mais recente disponível.

Dado que o HAI modela uma usina específica mista, foi preterida a detecção de ataques intra-processos em prol da detecção sistêmica, independente do processo associado ao ataque. Não só trata-se de uma abordagem mais geral como vai de encontro ao entendimento de que alguns ataques, apesar de não apresentarem efeitos nas demais variáveis intra-processuais, podem refletir nos outros processos, caracterizando uma anomalia no sistema.

Em relação à limpeza de dados, não houve valores nulos ou indisponíveis e não foram verificados dados de tipo incompatível com a sua *feature* de origem. Houve, contudo, atributos de valor constante, os quais foram descartados. Para a classe de interesse e de modo a padronizar os arquivos de entrada para a etapa de amostragem, foi considerada apenas a classe "attack", com foco na detecção sistêmica, em detrimento das classes "attack_P#", não presentes na versão de 2022 do HAI.

Apesar da disponibilidade específica de arquivos "train" (treino), dados de comportamento normal ou benigno existem também nos arquivos "test" (teste). Para diferenciar os casos originados da combinação de arquivos treino e teste, foi acrescentado um meta-atributo para indicar se o dado é originário de arquivo `train#.csv` ou não. A decisão tomada foi manter a finalidade desses registros nos casos presentes, de modo que, na etapa de construção dos conjuntos de treino e teste, tais instâncias estarão presentes apenas nos conjuntos de treino.

Não foi escopo da pesquisa aprofundar nas técnicas de seleção de atributos, tendo o entendimento que podem se estender para uma outra pesquisa e que podem ser adaptadas conforme a situação exigir, sem mudanças na proposta. Foram utilizados, portanto, dois métodos nativos do Scikit-learn para a seleção de *features*: `SelectFromModel` e `RFECV` (*Recursive Feature Elimination with Cross-Validation*).

Uma vez que os dados são amostrados para a construção dos conjuntos de treino e teste, perde-se a informação temporal dos registros, atributo descartado durante o pré-processamento. Para traduzir as informações temporais nos dados presentes, foi proposto um esquema de enriquecimento de *features* baseado em valores dos últimos atributos capturados, prévio à amostragem. Antes da seleção são computados, para cada atributo, os valores máximos e mínimos, as médias móveis simples e exponencial, e o desvio-padrão dos últimos n registros. Foram escolhidos os valores 5 e 10 para n, recordando que as capturas ocorrem a cada segundo. Dado que os cenários simulados de ataque ocorreram de forma contínua por um intervalo de tempo mínimo de 45 segundos, valores de n maiores podem camuflar os dados em situação de ataque com situação de normalidade. Por outro lado, valores muito pequenos exageram mudanças bruscas ou ruídos em uma sequência.

[2] HAI 20.07
[3] HAI 22.04

Neste trabalho, o termo “derivação” é utilizado para designar os *datasets* resultantes da combinação aleatória de arquivos treino e teste seguida ou da seleção de *features* ou do enriquecimento de *features* e posterior seleção pelo método *SelectFromModel*. Ressalta-se que, nesta etapa, os termos *train* e *test* referem-se aos nomes dos arquivos `.csv` disponíveis na pasta da versão do HAI, não se tratando dos habituais conjuntos treino-teste de Aprendizado de Máquinas, que serão vistos na próxima subseção.

4.2.3 Amostragem e construção dos conjuntos de treino e teste

Com base nas derivações, segue-se à amostragem dos dados. O tamanho da amostragem para cada derivação foi determinado pela média populacional, conforme a Equação 4. O valor da margem de erro E foi fixado em 0,01, grau de confiança bi-caudal de 90% e grau de liberdade tendendo ao infinito, resultando no valor da distribuição *TStudent* $Z\alpha/2$ em 1,645[4]. O valor de σ é determinado pelo maior desvio-padrão de cada *feature* normalizada, valor representativo da dispersão do universo amostral. Nos testes também foi utilizado o dobro do valor obtido na equação, para comparação.

$$n = \left(\frac{Z\alpha}{2} \times \frac{\sigma}{E} \right)^2 \tag{4}$$

Nesta etapa, são introduzidos parâmetros de execução para a criação de cenários a partir das derivações. Foram parametrizadas a proporção treino/teste e taxa de contaminação, i.e. proporção de dados de ataque. Os valores utilizados constam do Capítulo 5. Nas derivações formadas pela combinação de arquivo `train` e `test`, a construção dos conjuntos de treino obriga que os dados de comportamento normal sejam originados do arquivo treino. Pela construção do HAI, nos arquivos treino não há registros de ataque, de modo que, para permitir a classificação supervisionada os dados de ataque são retirados dos arquivos teste.

Tanto a amostragem quanto a construção dos conjuntos de treino e teste foram realizados sob a escolha de um parâmetro de randomização, para a reprodução de resultados. Soma-se a esse processo o aprendizado e predição, foram executadas 50 iterações para observar a distribuição dos resultados e minimizar ótimos locais.

4.2.4 Composição da detecção

A tarefa de classificação empregou os algoritmos do Scikit-learn constantes da tabela Tabela 2. Não foi escopo do trabalho o *tuning* dos classificadores, dado que almeja-se uma solução geral que possa ser especializada conforme a necessidade, e que a otimização, em muitos algoritmos de OD, é subjetiva e pode levar a resultados discrepantes (LI *et al.*, 2020).

[4] Extraído de https://pessoal.dainf.ct.utfpr.edu.br/maurofonseca.

Alguns classificadores agrupam outros e, para não ampliar por demais o escopo, foram restritos aos indicados na tabela. Os parâmetros de instanciação das classes foram os padrões.

A composição, por vezes referenciada neste trabalho como “duas etapas”, não necessita que os classificadores processem de forma serial. A estratégia empregada foi a soma das predições de cada registro, de modo que o dado será considerado um ataque caso pelo menos um dos dois classificadores assim o determine.

Tabela 2 – Classificadores empregados

Etapa	
Assinatura	Anomalia
Ada Boost	*One Class SVM - linear*
Bagging (Decision Tree)	*One Class SVM - radial basis function*
Bagging (Extra Trees)	*Local Outlier Factor - Novelty*
Bagging (Random Forest)	*Local Outlier Factor*
Decision Tree	*Isolation Forest*
Extra Trees	
Gaussian Naive Bayes	
Gaussian Process	
K-Nearest Neighbors (K=2)	
K-Nearest Neighbors (K=5)	
K-Nearest Neighbors (K=10)	
Logistic Regression	
Multilayer Perceptron	
Perceptron	
Quadratic Discriminant Analysis	
Random Forest	
Stochastic Gradient Descent	
Support Vector - linear	
Support Vector - radial basis function	

Os classificadores não supervisionados possuem um parâmetro de instanciação referente à taxa de contaminação do conjunto de dados. Nos experimentos foram executados tanto com o valor padrão quanto com o valor de taxa de contaminação herdado da taxa de contaminação da construção do cenário. Entende-se que, com a experiência e gestão do conhecimento pelo CSOC, esses valores podem ser utilizados e atualizados de sua base de dados.

4.2.5 Avaliação do desempenho

São avaliados dois conjuntos de predições: o proveniente da detecção por assinatura e o proveniente da detecção por composição. Das diversas métricas disponíveis na literatura, em um primeiro momento foi adotada a métrica *recall*, com amparo na motivação de dar maior importância a encontrar o maio número de anomalias, tendo em vista o impacto que podem gerar caso não detectadas. No decorrer na pesquisa, foram adotadas também as métricas acurácia,

F_1-*score* e AUC. Também são contadas as quantidades de FP e FN, e calculados FNR e FPR para a comparação entre cenários.

Foram gerados diagramas de distribuição em caixa para comparação entre classificadores supervisionados e entre classificadores não supervisionados, para cada cenário e métrica. Quanto à composição, é escolhido o par de classificadores de maior mediana da distribuição de escores da métrica avaliada. Quando, para o mesmo classificador supervisionado, há empate de classificadores não supervisionado, o critério de desempate é o menor número de falsos positivos gerados. O ganho de métricas é observado da comparação entre o valor da mediana da distribuição de escores pela composição e do obtido pela detecção por assinatura.

Para um mesmo cenário a contagem de falsos positivos e falsos negativos permite ao órgão decisor a opção por outras métricas. Conforme a opção por uma métrica ou outra, altera-se o par de classificadores em cada fase e repetidos experimentos de dados de mesma natureza podem entregar resultados que sirvam de guia de referência quando o CSOC obtiver informação suficiente para a construção de uma base de conhecimento. Entre cenários, a verificação de ambos FNR e FPR proporciona uma comparação normalizada para a obtenção de médias do comportamento das detecções em duas fases no âmbito das derivações. Espera-se a redução de FNR com o menor aumento de FPR, afinal, apesar do custo, é mais importante identificar e tratar os casos de ataques do que evitar falsos alarmes. Os resultados apresentados no Capítulo 5 permitem verificar a tendência com aplicação de diferentes métricas, bem como a possibilidade de ganhos pela detecção composta nas derivações do HAI.

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos pelo sistema proposto, enfatizando os ganhos percentuais nas métricas utilizadas, a comparação de falsos positivos e falsos negativos, os melhores e piores casos, os pares de classificadores mais recomendados e *insights* para uso do trabalho no futuro.

A detecção em duas etapas apresentou melhorias na métrica *recall* em comparação à detecção por assinatura para pelo menos um par de classificadores em 497 de 555 cenários, sendo que não houve, das vinte derivações do HAI, uma que não teve um cenário que não obteve benefício pela abordagem. Na Tabela 3, são apresentados os valores das médias dos ganhos percentuais mínimos, médios e máximos obtidos pela composição para a métrica *recall*.

Tabela 3 – Ganhos relativos em *recall* para derivações do HAI

Derivação	Mínimo	Médio	Máximo
hai20_te2_ext-sfm	0,697	11,24	70,26
hai20_te2_rfcv10	3,117	25,36	233,01
hai20_te2_sfm	2,144	19,70	88,71
hai20_tr1_te1_ext-sfm	0,003	4,16	61,12
hai20_tr1_te1_rfcv10	0,003	12,04	112,13
hai20_tr1_te1_sfm	0,006	11,16	98,06
hai22_te1_sfm	0,022	44,72	574,83
hai22_te4_rfcv10	0,033	156,30	1829,35
hai22_te4_sfm	0,011	394,41	19588,24
hai22_tr1_te1_ext-sfm	0,006	2,78	242,07
hai22_tr1_te1_rfcv10	0,006	299,45	66020,00
hai22_tr1_te1_sfm	0,005	47,99	1527,70
hai22_tr1_te2_ext-sfm	0,002	0,26	18,34
hai22_tr1_te2_rfcv10	0,042	2692,47	431300,00
hai22_tr1_te2_sfm	0,002	6,15	68,25
hai22_tr2_te1_ext-sfm	0,006	0,05	6,09
hai22_tr2_te1_rfcv10	0,005	3,62	125,29
hai22_tr2_te1_sfm	0,005	8,79	274,34
hai22_tr2_te4_sfm	0,004	1256,41	361950,00
hai22_tr4_te3_ext-sfm	0,001	0,03	4,71

Fonte: Autoria própria (2024).

Foi verificado que a métrica *recall* prioriza a completude das detecções, porém sem conter-se no número de falsos positivos. De fato, a simples opção por tratar todos os registros como anomalias garante *recall* em 1.0, mas por óbvio não reflete uma boa detecção. A métrica F_1*-score* contrabalanceia valores elevados de *recall*, levando em conta também a quantidade de falsos positivos. A detecção composta, para essa métrica, apresentou melhoria em comparação à detecção por assinatura em todas as derivações, com menor valor comparativo percentual inferior a 0.0001%.

Quanto à métrica *F_1-score*, a detecção em duas etapas apresentou melhorias na comparação à detecção por assinatura para pelo menos um par de classificadores em 453 de 555 cenários, sendo que não houve, das vinte derivações do HAI, uma que não teve um cenário que não obteve benefício pela abordagem. Na Tabela 4 são apresentados os valores das médias dos ganhos percentuais mínimos, médios e máximos obtidos pela composição para a métrica *F_1-score*.

Tabela 4 – Ganhos relativos em *F_1-score* para derivações do HAI

Derivação	Mínimo	Médio	Máximo
hai20_te2_ext-sfm	0,0002	1,99	22,22
hai20_te2_rfcv10	0,0018	7,77	112,63
hai20_te2_sfm	0,0001	4,88	27,76
hai20_tr1_te1_ext-sfm	0,0019	2,97	36,83
hai20_tr1_te1_rfcv10	0,0020	7,98	63,02
hai20_tr1_te1_sfm	0,0040	7,70	56,29
hai22_te1_sfm	0,0784	14,05	104,36
hai22_te4_rfcv10	0,0001	44,03	475,42
hai22_te4_sfm	0,0014	145,59	4908,70
hai22_tr1_te1_ext-sfm	0,0205	19,53	113,41
hai22_tr1_te1_rfcv10	0,7028	222,18	22960,78
hai22_tr1_te1_sfm	0,0012	19,05	397,27
hai22_tr1_te2_ext-sfm	0,0017	0,10	0,90
hai22_tr1_te2_rfcv10	0,0008	896,54	104509,51
hai22_tr1_te2_sfm	0,0027	1,65	8,04
hai22_tr2_te1_ext-sfm	0,0052	0,04	0,07
hai22_tr2_te1_rfcv10	0,0034	3,63	17,22
hai22_tr2_te1_sfm	0,0086	11,72	41,49
hai22_tr2_te4_sfm	0,0004	324,07	70318,17
hai22_tr4_te3_ext-sfm	0,0005	0,09	0,84

Fonte: Autoria própria (2024).

Na métrica AUC, também houve melhoria pela composição quando comparada à detecção por assinatura em todas as derivações, com 462 de 555 cenários com ao menos um par com ganho de métrica. Percebem-se, na Tabela 5, ganhos máximos mais contidos quando comparados aos máximos recentemente vistos.

Por fim, a Tabela 6 mostra os ganhos médios para a métrica acurácia, a única que deixou de apresentar melhorias em duas das vinte derivações. Dos 555 cenários, 387 apresentaram melhorias pela composição quando comparados à detecção por assinatura. Os ganhos são menores quando comparados às demais métricas e, observando os valores absolutos foi verificado que para as duas derivações em que não houve melhorias foram alcançados valores de acurácia em 0,966 e 0,964, com os mesmos valores para a composição.

Valores elevados de ganho relativo não significam que o desempenho final da detecção foi próximo a 100%. De fato, por construção, quanto mais próximo de 1,00 for a avaliação de métrica da primeira fase, menor o espaço para melhorias pela composição e, portanto, menor

Tabela 5 – Ganhos relativos em AUC para derivações do HAI

Derivação	Mínimo	Médio	Máximo
hai20_te2_ext-sfm	0,0012	5,04	93,15
hai20_te2_rfcv10	0,0053	6,96	48,16
hai20_te2_sfm	< 0,0001	5,06	25,13
hai20_tr1_te1_ext-sfm	0,0029	3,80	68,26
hai20_tr1_te1_rfcv10	0,0030	6,39	34,07
hai20_tr1_te1_sfm	0,0060	6,32	31,71
hai22_te1_sfm	0,0159	5,14	24,57
hai22_te4_rfcv10	0,0009	7,96	42,99
hai22_te4_sfm	0,0023	11,77	52,85
hai22_tr1_te1_ext-sfm	0,0054	11,06	39,96
hai22_tr1_te1_rfcv10	0,0980	14,19	31,47
hai22_tr1_te1_sfm	0,0023	8,85	38,13
hai22_tr1_te2_ext-sfm	0,0006	1,42	26,89
hai22_tr1_te2_rfcv10	0,0043	9,67	26,25
hai22_tr1_te2_sfm	< 0,0001	1,43	7,29
hai22_tr2_te1_ext-sfm	0,0131	23,53	47,31
hai22_tr2_te1_rfcv10	0,0054	2,97	15,90
hai22_tr2_te1_sfm	0,0012	6,32	20,53
hai22_tr2_te4_sfm	0,0013	7,54	32,89
hai22_tr4_te3_ext-sfm	0,0003	0,08	0,65

Fonte: Autoria própria (2024).

o ganho relativo. Os maiores resultados médios de ganhos relativos de métricas entregaram valores absolutos de métricas de até 0,52 para acurácia, 1.0 em *recall*, 0,94 para F_1*-score* e 0,95 em AUC. Já os menores ganhos médios entregaram até 0,93 de acurácia, 1,0 em *recall*, F_1*-score* em 0,69 e AUC de 0,64. Isso ilustra que não necessariamente há a correlação entre ganhos e métricas.

As detecções mostraram-se bem dependentes das derivações, isto é, cenários cujas métricas são de baixo valor tendem a ser oriundos da mesma derivação. A classificação não supervisionada acompanhou o desempenho da supervisionada, indicando que cada derivação tratou de conjunto de dados particulares e independentes do método utilizado. Ainda nesses casos em que a detecção por assinatura não desempenhou bem, foi possível verificar melhoria com a abordagem em duas etapas.

A Figura 6 apresenta o gráfico de melhores métricas pela detecção por assinatura de cada cenário que apresentou melhoria, em azul, e sua métrica correspondente para a detecção em duas fases, na cor laranja. Os pares foram ordenados primeiramente pelo valor obtido pela composição, seguido do valor obtido pela assinatura. É possível visualizar para as métricas *recall*, F_1*-score* e AUC a existência de maiores ganhos para valores menores de métrica, como esperado. A métrica acurácia foi mais contida nos ganhos, corroborando com os dados da Tabela 6.

Tabela 6 – Ganhos relativos em acurácia para derivações do HAI

Derivação	Mínimo	Médio	Máximo
hai20_te2_ext-sfm	< 0,0001	0,51	9,56
hai20_te2_rfcv10	< 0,0001	1,59	11,01
hai20_te2_sfm	< 0,0001	1,10	6,53
hai20_tr1_te1_ext-sfm	0,0029	0,95	5,84
hai20_tr1_te1_rfcv10	0,0030	2,35	9,40
hai20_tr1_te1_sfm	0,0060	2,45	9,50
hai22_te1_sfm	< 0,0001	0,44	2,13
hai22_te4_rfcv10	< 0,0001	0,30	15,68
hai22_te4_sfm	< 0,0001	0,61	8,72
hai22_tr1_te1_ext-sfm	0,9537	1,92	3,50
hai22_tr1_te1_rfcv10	-	-	-
hai22_tr1_te1_sfm	< 0,0001	1,55	6,83
hai22_tr1_te2_ext-sfm	0,0079	0,05	0,15
hai22_tr1_te2_rfcv10	< 0,0001	1,87	12,54
hai22_tr1_te2_sfm	< 0,0001	0,14	0,66
hai22_tr2_te1_ext-sfm	0,0399	0,06	0,10
hai22_tr2_te1_rfcv10	-	-	-
hai22_tr2_te1_sfm	0,0082	0,01	0,01
hai22_tr2_te4_sfm	< 0,0001	1,04	9,76
hai22_tr4_te3_ext-sfm	0,0059	0,02	0,03

Fonte: Autoria própria (2024).

Na Figura 7 é apresentado um dos melhores casos encontrados, em que a composição proposta não apenas trouxe ganhos em 33,33% na métrica *recall* (quase 0,2 de ganho absoluto), mas principalmente apresentou duas informações: o incremento de uma unidade de falso positivo e a zeragem dos falsos negativos. Neste cenário, foi possível verificar que a detecção por assinatura por si só já resulta em um bom desempenho para esse conjunto de dados, porém a detecção composta permitiu, com um baixo incremento no número de falsos positivos, encontrar os ataques faltantes.

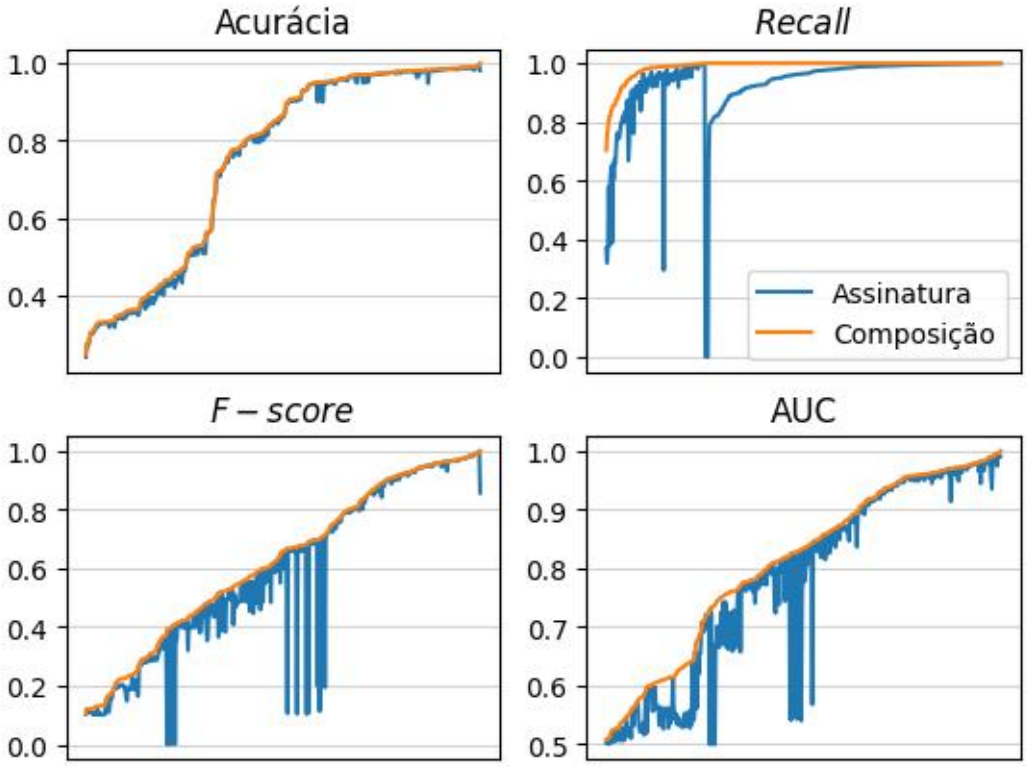
Tabela 7 – Efeito da composição para FP e FN

Métrica	Aumento médio FP	Redução média FN	Aumento médio FPR	Redução média FNR
Acurácia	8,84	16,76	1,19%	8,19%
Recall	454,67	34,28	36,17%	19,36%
F_1*-score*	48,32	32,38	4,48%	14,03%
AUC	49,53	27,06	4,13%	13,66%

Fonte: Autoria própria (2024).

Como a detecção por assinatura não apresentou falsos positivos para a maioria dos classificadores ilustrados, não houve esforço desperdiçado pela equipe de tratamento de incidentes: todos os registros detectados como anomalias de fato o eram, porém nem todas anomalias foram encontradas. Já com a composição, o incremento de um falso positivo, em termos

Figura 6 – Comparação de métricas por detecção.



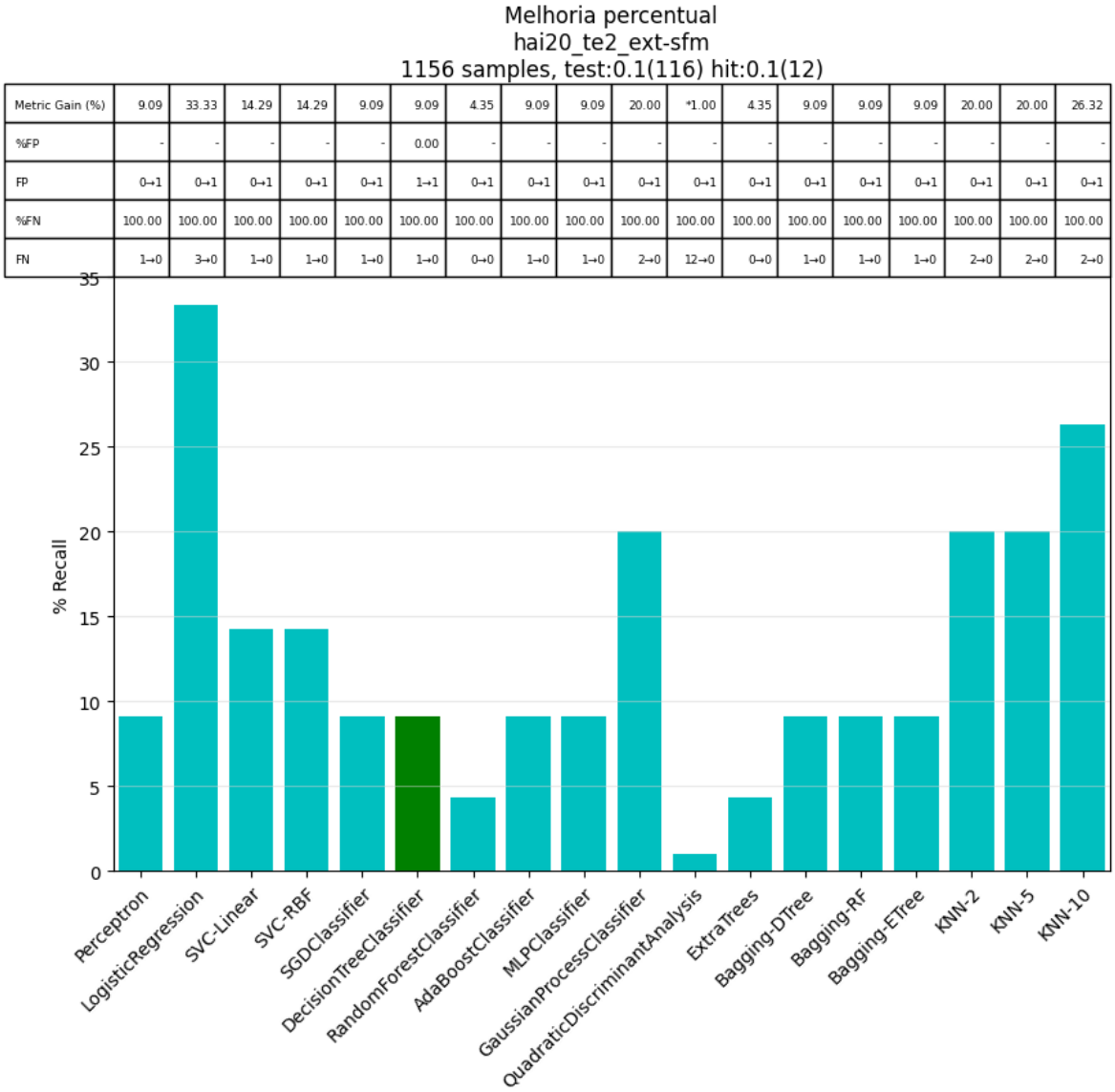
Fonte: Autoria própria (2024).

de desperdício de esforço humano para a análise de um registro que no fim era normal, pode compensar pelo fato de outra anomalia ter sido encontrada. No caso em questão, a anomalia que não foi detectada pela abordagem por assinatura o foi pela composição de classificadores supervisionados e não supervisionados, ao custo da análise pelo CSOC de mais dois registros: um falso positivo e um verdadeiro positivo.

As melhorias obtidas quanto ao aumento da detecção de anomalias é refletido na métrica *recall*, porém não foi incomum observar o aumento expressivo do número de falso positivos, ainda que, por vezes, com a zeragem dos falsos negativos. Nesses casos, a avaliação pelas métricas F_1*-score* e AUC vêm a oferecer soluções intermediárias para o elemento decisor da organização responsável pela infraestrutura crítica. A Tabela 7 apresenta o aumento médio de falsos positivos, para todos os cenários em que houve melhoria pela composição, e a redução média de falsos negativos, comparados à detecção por assinatura. Como os cenários possuem quantidades distintas de ataques, são apresentadas também as taxas relativas de falsos positivos e falsos negativos.

Na comparação, percebe-se que, enquanto a composição avaliada pela métrica *recall* apresenta a maior redução de falsos negativos (e portanto detecta mais ataques), o custo para o CSOC, na forma de falsos positivos, é elevado (e portanto há maior dispêndio de recurso humano para tratar os falsos alarmes). A métrica F_1*-score* traz uma alternativa para o órgão decisor: com uma redução de falsos negativos um pouco menor quando comparada à mé-

Figura 7 – Ganho percentual em *recall* por classificador supervisionado.



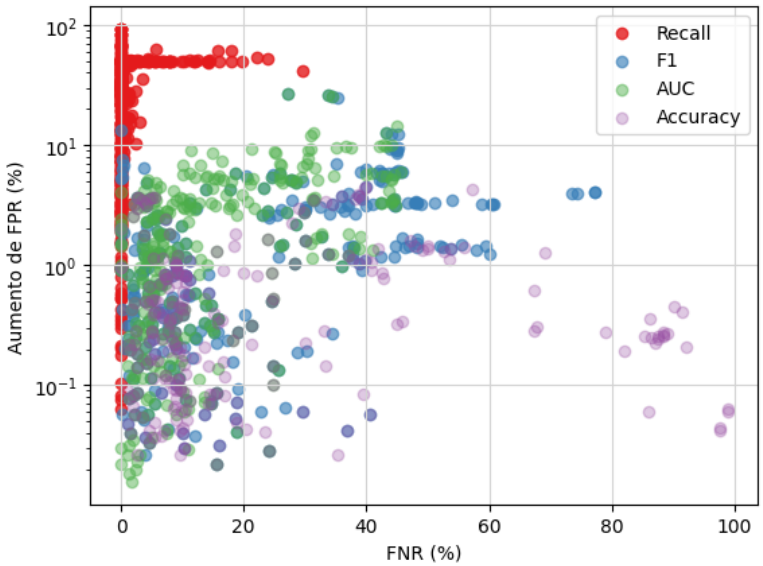
Fonte: Autoria própria (2024).

trica *recall* (de 19,36% para 14,03%), o aumento de falsos positivos é expressivamente menor (36,17% contra 4,48% de FPR).

A Figura 8 corrobora com as médias encontradas na Tabela 7, permitindo identificar o comportamento da composição segundo cada métrica quanto ao FNR final e aumento percentual de FPR. A métrica *recall* destaca-se tanto pelo menor FNR final quanto pelo maior aumento de FPR.

Em alguns cenários, a métrica AUC apresentou-se como meio-termo quando comparados os resultados de um mesmo cenário entre as métricas F_1*-score* e *recall*. Ou seja, se da análise da composição com base na métrica F_1*-score* o elemento decisor desejar aumentar a detecção das anomalias sem aumentar expressivamente o dispêndio de recurso humano no tratamento de falsos positivos, poderá optar pela composição com avaliação pela métrica AUC.

Figura 8 – Dispersão das melhores detecções por métrica.



Fonte: Autoria própria (2024).

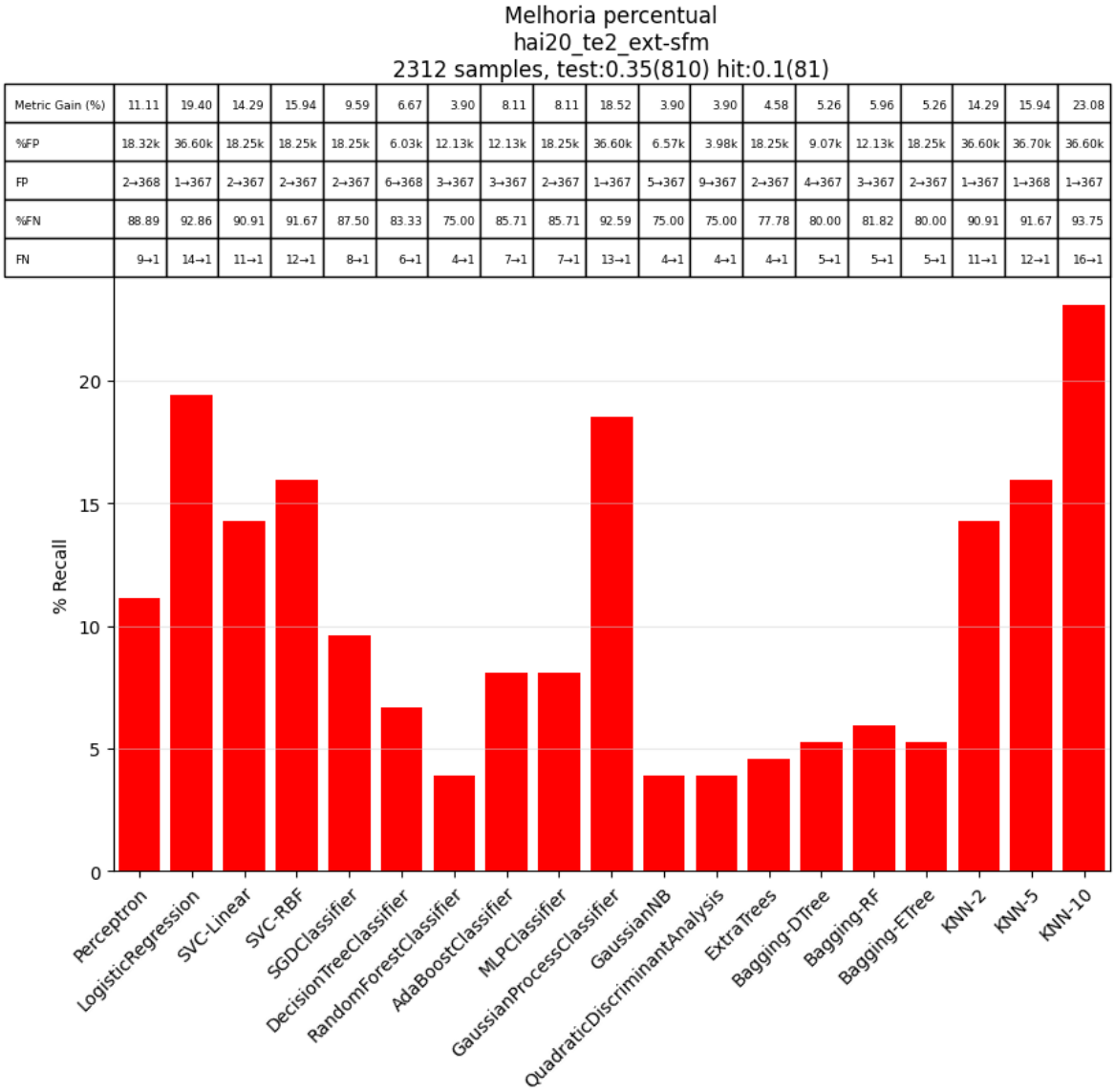
A Figura 9 apresenta um cenário em que a detecção em duas etapas, quando comparada com a detecção por assinatura, trouxe aumentos variando de 3.9% a 23% na métrica *Recall* (0,0375 a 0,18 de variação em termos absolutos), contudo a um custo elevado em termos de falsos positivos. Mesmo encontrando mais anomalias, nesse cenário restou um registro sem ser detectado e houve um aumento de mais de 360 registros classificados equivocadamente como anomalias. Por certo que o impacto do aumento no número de falsos positivos é menor em termos de danos do que o aumento ou estagnação no número de falsos negativos, porém caso a política da organização exija um menor número de falsos alarmes, a adoção de outra composição pode satisfazer essa necessidade.

No cenário em questão, o classificador KNN-10 encontrou 65 das 81 anomalias e classificou erroneamente um registro normal. Se a adoção for pela métrica *Recall*, são encontradas outras 15 anomalias não detectadas anteriormente, porém os registros de atividade normal que serão analisados em vão pela equipe de tratamento de incidentes sobe em 366 unidades, quase 45% do conjunto de teste. A composição, nesse caso, deu-se pelo par KNN-10 e OneClassSVM-RBF, como observado na Figura 10.

A Figura 11 apresenta o mesmo cenário sob a ótica da métrica F_1*-score*. Pode-se observar que um número menor de classificadores supervisionados teve um aumento na distribuição de métricas, pela construção do gráfico. Em ambos os casos o KNN-10 teve o maior aumento percentual na comparação entre a detecção em duas etapas e por assinatura. A opção pela métrica F_1*-score* resultou, comparado à *Recall*, um menor número de anomalias encontradas fora da assinatura: 9 contra 15. Por outro lado, o aumento de falsos positivos foi de apenas três unidades contra os 366 vistos na Figura 10.

Essa diferença tornar-se-á ainda mais expressiva quanto mais tempo a equipe de tratamento de incidentes levar para avaliar manualmente cada registro. Se o elemento decisor pon-

Figura 9 – Ganho percentual em *recall* por classificador supervisionado.



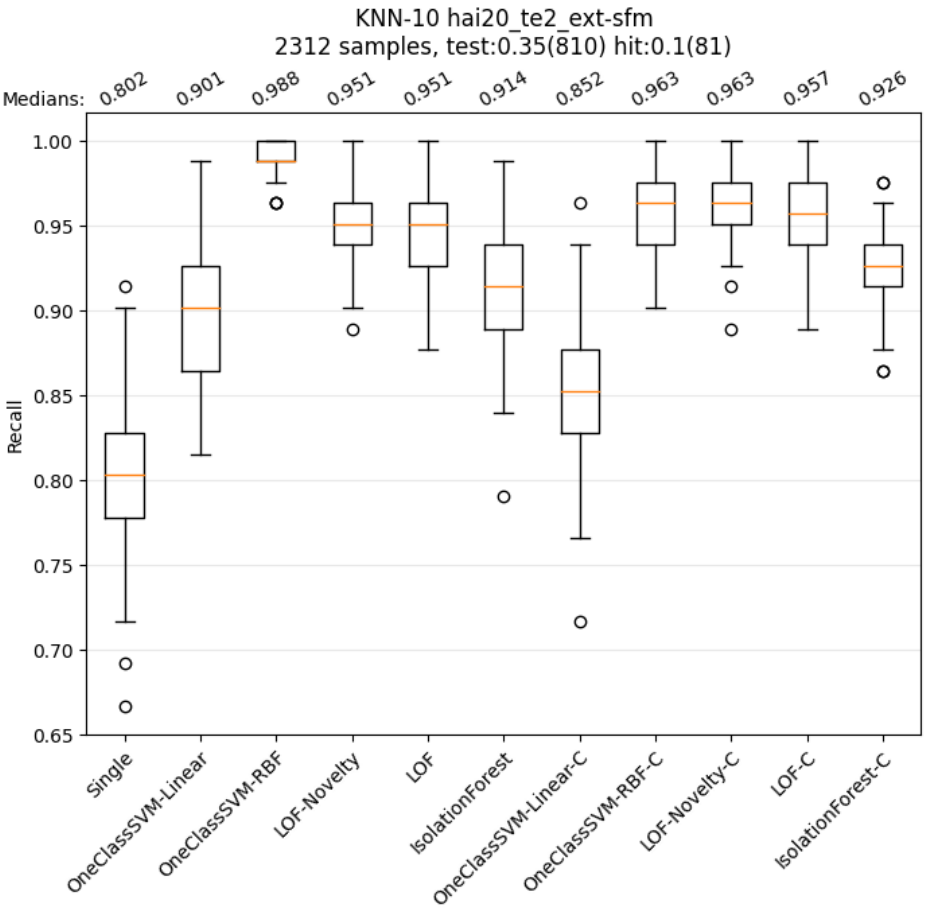
Metric Gain (%)	11.11	19.40	14.29	15.94	9.59	6.67	3.90	8.11	8.11	18.52	3.90	3.90	4.58	5.26	5.96	5.26	14.29	15.94	23.08
%FP	18.32k	36.60k	18.25k	18.25k	18.25k	6.03k	12.13k	12.13k	18.25k	36.60k	6.57k	3.98k	18.25k	9.07k	12.13k	18.25k	36.60k	36.70k	36.60k
FP	2→368	1→367	2→367	2→367	2→367	6→368	3→367	3→367	2→367	1→367	5→367	9→367	2→367	4→367	3→367	2→367	1→367	1→368	1→367
%FN	88.89	92.86	90.91	91.67	87.50	83.33	75.00	85.71	85.71	92.59	75.00	75.00	77.78	80.00	81.82	80.00	90.91	91.67	93.75
FN	9→1	14→1	11→1	12→1	8→1	6→1	4→1	7→1	7→1	13→1	4→1	4→1	4→1	5→1	5→1	5→1	11→1	12→1	16→1

Fonte: Autoria própria (2024).

derar pela aceitação do risco em deixar passar as anomalias não detectadas, para esse cenário poderá adotar a composição entre o *KNN-10* e o *Isolation Forest*, como pode ser observado na Figura 13.

Apesar do aumento médio relativo de ganho de métricas, alguns classificadores supervisionados entregam métricas muito menores quando comparados aos demais, o que contribui para inflar o aumento relativo do ganho de métrica pela composição. Ainda assim, adotando um limiar de 0,9 para a detecção por assinatura para depois considerar os ganhos de métrica, são obtidos ganho médio relativo de 3,76% para a métrica *recall*, 0,68% para acurácia, 0,63% de F_1*-score* e 1,3% em AUC. Pode parecer insignificante em termos numéricos mas na semântica pode fazer muita diferença quando colocado em perspectivas os danos evitados de um ataque que passou a ser detectado. A Figura 12 ilustra o ganho absoluto (cor laranja) nos cenários em

Figura 10 – Distribuição em *recall* para composição com KNN-10.



Fonte: Autoria própria (2024).

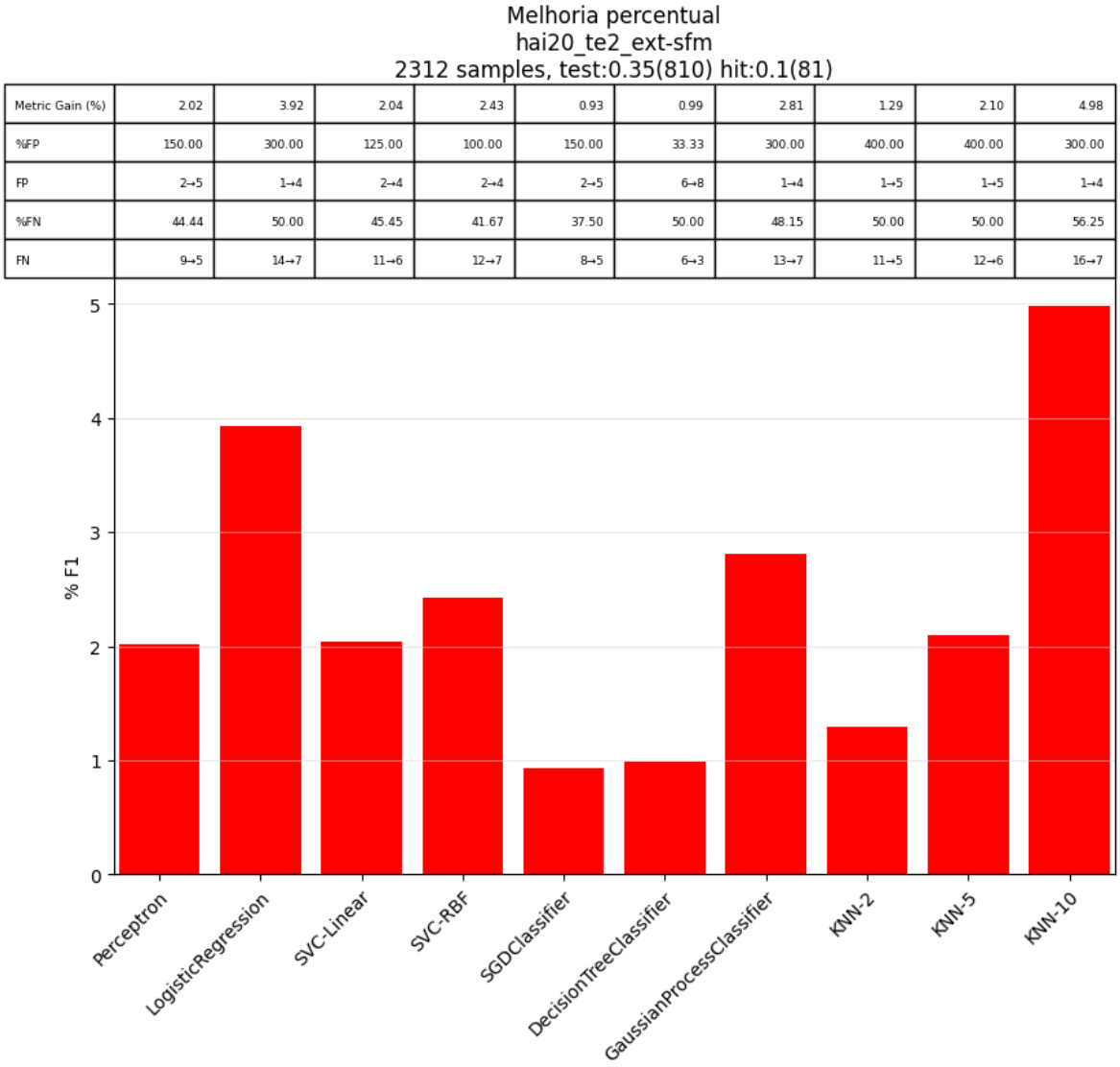
que foi aplicado o limiar na detecção por assinatura, cujo valor de métrica é representado em azul.

Dentre os classificadores supervisionados, o *ExtraTrees* esteve entre os melhores avaliados em todas as métricas. A Figura 14 permite visualizar a comparação dos classificadores supervisionados que tiveram melhor desempenho por cenário em cada métrica. A análise continuada dos melhores classificadores pode amparar o descarte dos demais para redução de experimentos e a designação condicionada à métrica de interesse.

Para os classificadores não supervisionados o *Local Outlier Factor*, considerando tanto o modo detecção de novidade quanto OD e parâmetro `contamination` padrão ou definido (sufixo “-C"), à exceção da métrica *recall*, representam mais de 50% dos melhores avaliados para a detecção composta. A Figura 15 permite verificar que o *Isolation Forest* também sobressaiu os demais e a avaliação pela métrica *recall* apresenta ordenamento mais distribuído quando comparada às demais.

Em relação aos pares de classificadores não foi encontrado um par que se destaque sobremaneira em relação aos demais. A Tabela 8 apresenta apenas os três primeiros pares para cada métrica em ordem decrescente de quantidade de cenários em que o par obteve as melhores métricas. Em caso de empates, os critérios foram o menor número de falsos negati-

Figura 11 – Ganho percentual em *F_1-score* por classificador supervisionado.



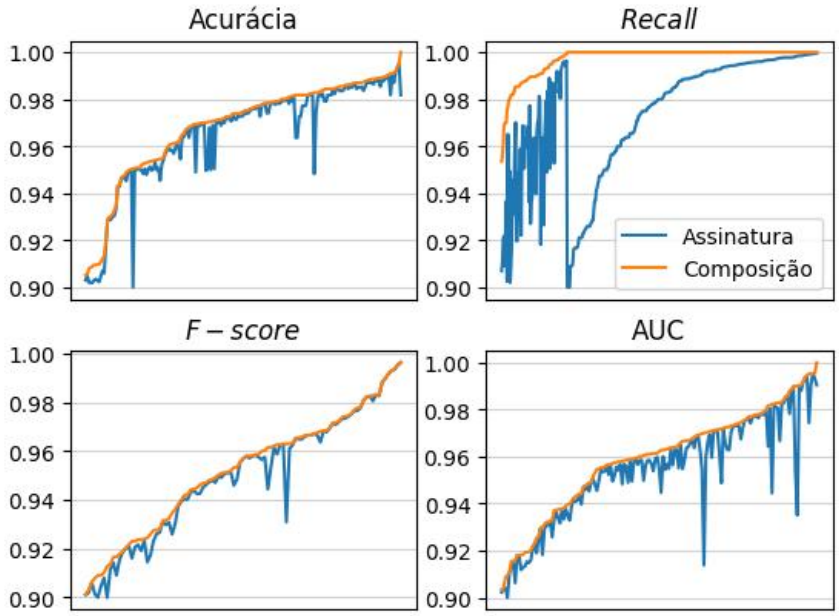
Metric Gain (%)	2.02	3.92	2.04	2.43	0.93	0.99	2.81	1.29	2.10	4.98
%FP	150.00	300.00	125.00	100.00	150.00	33.33	300.00	400.00	400.00	300.00
FP	2→5	1→4	2→4	2→4	2→5	6→8	1→4	1→5	1→5	1→4
%FN	44.44	50.00	45.45	41.67	37.50	50.00	48.15	50.00	50.00	56.25
FN	9→5	14→7	11→6	12→7	8→5	6→3	13→7	11→5	12→6	16→7

Fonte: Autoria própria (2024).

vos e, em seguida, de falsos positivos. Ao serem comparados com o desempenho isolado dos classificadores supervisionados depara-se com uma aparente inconsistência, dado que, pela Figura 14, os classificadores `ExtraTrees`, `KNN-2` e `GaussianProcessClassifier` são os de maior contagem de cenários. Contudo, vale lembrar que não somente alguns cenários não tiveram melhorias pela composição e que a composição é avaliada pelo resultado composto da métrica e não pela média ou soma das métricas de cada etapa. De fato, para as métricas acurácia, *F_1-score* e AUC, não se beneficiaram da composição, respectivamente, 168, 102 e 93 cenários. Nesses casos, somente para a métrica *F_1-score* mais de 50% desse cenários obtiveram valores inferiores a 0,5.

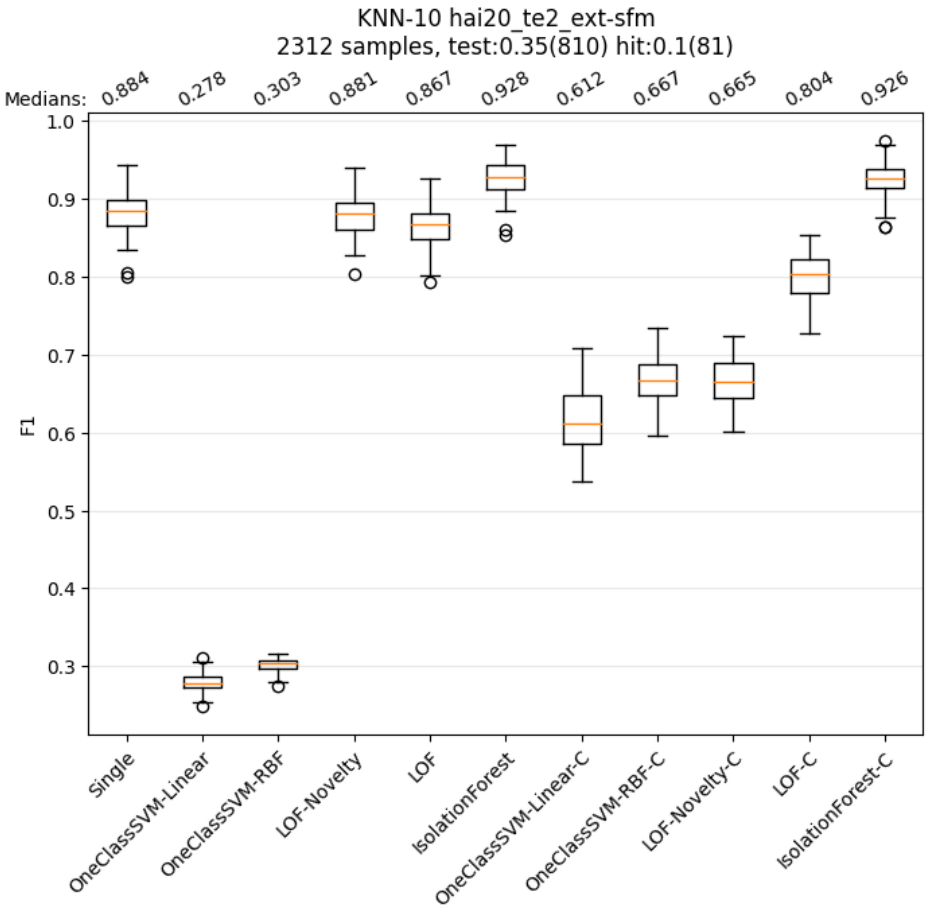
Quanto à verificação da construção dos cenários foi verificado que, no âmbito das derivações, quando a amostragem se deu pelo valor original extraído da Equação 4, os resultados

Figura 12 – Comparação de métrica por detecção com restrição de limiar.



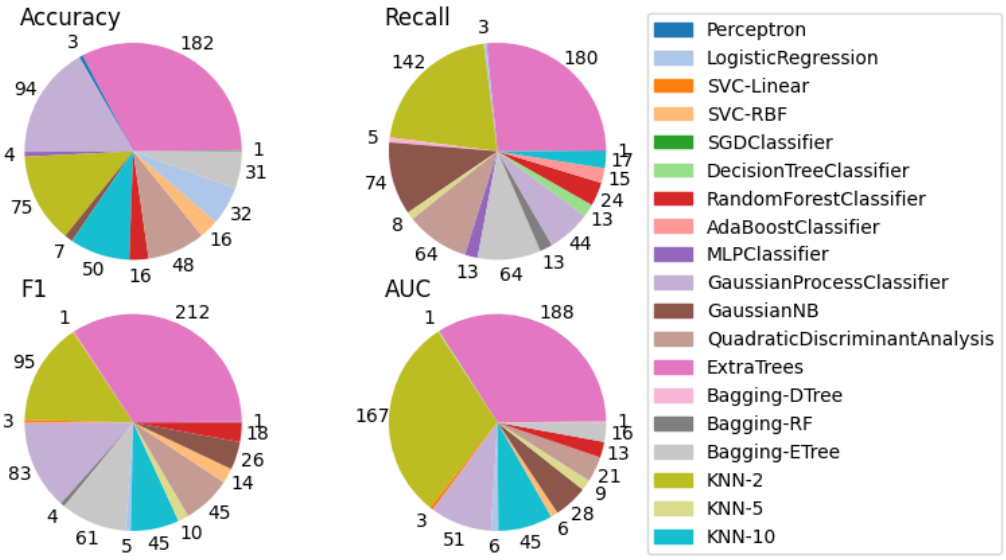
Fonte: Autoria própria (2024).

Figura 13 – Distribuição em F_1-*score* para composição com KNN-10.



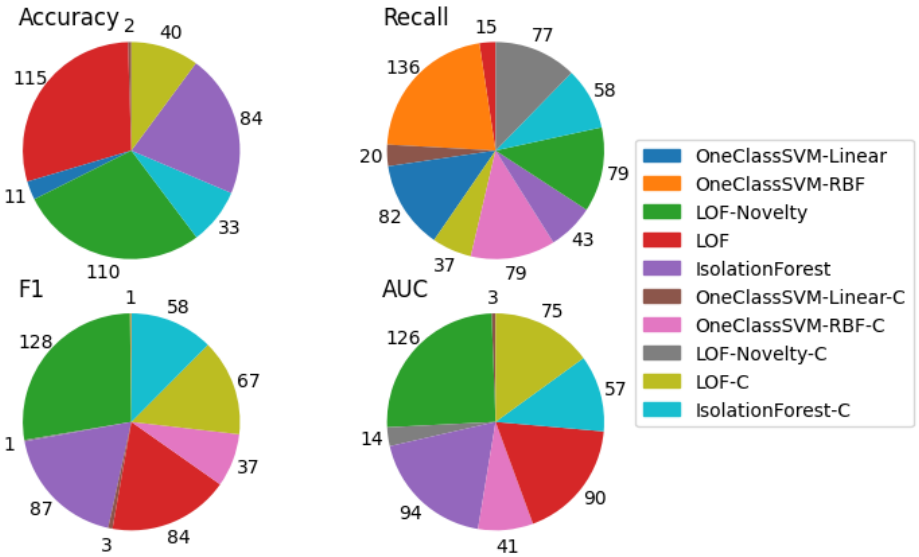
Fonte: Autoria própria (2024).

Figura 14 – Contagem de cenários para melhor classificador supervisionado



Fonte: Autoria própria (2024).

Figura 15 – Contagem de cenários para melhor classificador não supervisionado



Fonte: Autoria própria (2024).

Tabela 8 – Contagem de cenários para os três primeiros pares

Métrica	Par		Quantidade
	Assinatura	Anomalia	
Acurácia	QuadraticDiscriminantAnalysis	LOF	28
	LogisticRegression	LOF-C	22
	LogisticRegression	LOF	19
Recall	LogisticRegression	OneClassSVM-RBF	36
	ExtraTrees	OneClassSVM-Linear	36
	LogisticRegression	LOF-Novelty	26
F_1-*score*	LogisticRegression	LOF-C	30
	KNN-2	LOF-Novelty	21
	Bagging-ETree	LOF	19
AUC	LogisticRegression	LOF-C	37
	GaussianProcessClassifier	IsolationForest	19
	KNN-10	LOF	18

Fonte: Autoria própria (2024).

tenderam a ser superiores em 72,66% dos casos quando comparados ao uso da amostragem dobrada.

Para o parâmetro de partição em conjuntos de treino e teste, para as métricas acurácia e *F_1-score* a proporção 0,5 entre treino e teste resultou em melhores detecções em 56,4% e 44,9% dos casos, respectivamente, quando comparada às proporções de 0,65 para 0,35 e de 0,9 para 0,1. Para as métricas *recall* e AUC, a proporção de 0,9 de treino e 0,1 de teste apresentou melhores detecções em 92% e 51,02%, respectivamente. Em todas as métricas a partição de 0,35 para o conjunto de teste foi a que teve menor quantidade de cenários com maior detecção quando comparada às demais proporções.

O último parâmetro para a criação de cenários é a taxa de contaminação, isto é, a proporção de ataques presentes no conjunto de treino e teste, para o aprendizado supervisionado, e ataques presentes apenas no conjunto de teste para o aprendizado não supervisionado. Os valores padrão são 0,05, 0,1, 0,25 e 0,5. Nos casos em que a derivação não possuía registros de ataques suficientes para construção do cenário, foi construído um cenário especial com a máxima taxa permitida pela limitação do espaço amostral. O valor padrão de 0,05 resultou na maior quantidade de melhores detecções em três das quatro métricas, na média de 47,17% em comparação aos cenários dos demais valores. Para a métrica *F_1-score* o valor padrão 0,5 gerou melhores detecções em 45,71% dos casos. Ademais, na métrica *recall*, resultou em 33,78% das detecções, muito próximo ao obtido pelo valor 0,05, em 35,14%.

Em relação à proporção de melhores detecções geradas pela taxa de contaminação em 0,5, o resultado corrobora com a expectativa de que um número maior de exemplos de ataques resulta em melhor aprendizado e desempenho pela detecção por assinatura, ainda mais na composição com o aprendizado não supervisionado. Por outro lado, o valor 0,05 aproxima-se de uma situação real, em que há a predominância de instâncias de comportamento normal.

Esse valor provoca o desbalanceamento de classes e consequente tendência de aumento nas taxas de acurácia.

Para resumir, a detecção em duas fases pela composição de classificação supervisionada e não supervisionada apresentou ganhos relativos de métrica na ordem de 32,30% para a métrica AUC, 6,187% para a métrica acurácia, 59,658% para a métrica F_1-*score* e 179,15% para a métrica *recall*. Em termos de ganhos absolutos, esses valores apresentam, respectivamente, ganhos de 0,161, 0,044, 0,143 e 0,687. Dado que classificações de menor desempenho podem inflar os ganhos relativos, quando aplica-se um limiar de 0,9 para selecionar os cenários cujas detecções por assinatura sejam superiores a esse limiar, ainda assim são observados ganhos na ordem de 3,76% para a métrica *recall*, de 0,68% para acurácia, de 0,63% de F_1-*score* e 1,3% em AUC. Nesses casos, foi possível alcançar, com a composição, métricas na média de 0,968 para acurácia, 0,998 para *recall*, 0,952 em F_1-*score* e 0,959 para AUC. A redução de FNR com a composição foi maior quando avaliada sob a ótica da métrica *recall*, em 19,36%, seguida da F_1-*score* e AUC. O aumento médio de FPR foi de 36,17% sob a métrica *recall* e abaixo de 4,48% para as demais métricas, ofertando possibilidades de escolha para o elemento decisor responsável pela infraestrutura crítica quanto ao *trade-off* entre detecção e falsos alarmes. Não houve um par de classificadores que se destacou dos demais, no entanto os classificadores *ExtraTrees* e *KNN-2* tiveram melhor desempenho na detecção por assinatura que os demais supervisionados, enquanto que para os não supervisionados o *Local Outlier Factor* e suas variações estiveram presentes na maioria dos melhores desempenhos por cenário.

6 CONCLUSÃO

Neste trabalho, foi adotado o HAI, um recente *dataset* obtido pelo emprego de um simulador *hardware-in-the-loop* e componentes reais, para a detecção de anomalias causadas por ataques FDI em uma infraestrutura crítica representada por uma usina de geração de energia elétrica. Nos trabalhos relacionados ao HAI foi verificada uma forte preferência pela pesquisa no uso de *Deep Learning*, oferecendo uma lacuna para a pesquisa de detecção de ataques com métodos tradicionais de Aprendizado de Máquinas. Foi proposta uma detecção composta por classificadores supervisionados e não supervisionados, para obter ambas vantagens da detecção por assinatura e por anomalia. Foram criadas vinte derivações e 555 cenários a partir dos arquivos do HAI. Os testes foram executados em 50 iterações com parâmetro de aleatorização controlado. Além da comumente empregada acurácia, as detecções foram avaliadas sob a ótica das métricas *recall*, F_1*-score* e AUC.

A métrica *recall*, em especial, oferece uma melhor completude das detecções, porém a um custo de falsos alarmes maior que seus pares. Nesse sentido, foi verificada a possibilidade da opção pela métrica F_1*-score* e AUC, com bom aumento das detecções pela abordagem composta quando comparado à detecção por assinatura, sem um elevado aumento de falsos positivos.

Foi verificado que em pelo menos 89,55% dos cenários a detecção composta é capaz de melhorar o desempenho da avaliação, variando de 32,3% a 179,15% de ganho relativo médio conforme a métrica. Se forem considerados apenas os casos em que a fase da detecção por assinatura obtém métricas acima de 0,9, os ganhos naturalmente diminuem pela proximidade com o valor de 100%, no entanto são observados ganhos médios relativos variando de 0,63% a 3,76% conforme a métrica. Considerando a premissa do tratamento dos incidentes pela equipe do CSOC, a detecção composta apresentou situações em que o aumento da métrica, apesar do incremento de falsos alarmes, resultou na redução total de falsos negativos e portanto todos ataques do cenário foram detectados. Por outro lado, foi verificado um relativo compromisso entre a redução de falsos negativos e aumento de falsos positivos. Para a métrica *recall*, a composição apresentou em média redução de 19,36% de FNR com aumento de FPR em 36,17%. Nos casos em que seja possível flexibilizar a detecção para reduzir os falsos alarmes, a composição avaliada pela métrica F_1*-score* apresentou redução de FNR pouco menor, na ordem de 14,03%, porém com expressivo menor aumento de FPR em comparação ao obtido pela métrica *recall*, na média em 4,48%.

Em termos absolutos, nos melhores casos foi possível obter com a composição, na média, métricas valorando 0,968 para acurácia, 0,998 em *recall*, 0,952 para F_1*-score* e 0,959 de AUC. Ademais, foi possível verificar que, sob a métrica *recall*, em todas as vinte derivações ocorreram cenários em que a composição encontrou todos os ataques sem ingenuamente acusar todos os registros como tal. Já para as métricas acurácia, F_1*-score* e AUC isso aconteceu em sete, nove e dez derivações, respectivamente.

Em termos de algoritmos, verificou-se que os baseados em árvore destacam-se em relação aos demais no aprendizado supervisionado. Para a detecção de anomalias, o *Local Outlier Factor* e suas variações destacaram-se para quase todas as métricas, à exceção da *recall*, que teve os melhores resultados com o *One Class Support Vector Machine*. Quanto aos pares de classificadores da composição, não houve um que se destacou sobremaneira frente aos demais.

Para trabalhos futuros vislumbra-se a construção de uma plataforma de testes para captura de dados de ICS simulada em HIL baseada em infraestruturas críticas reais, como a Itaipu, para permitir a construção de cenários de ataques cibernéticos e fornecer dados de pesquisa para detecção de ataques. Apesar de o *dataset* do HAI permitir amplas oportunidades de pesquisa com novas métricas e algoritmos, ter um *dataset* baseado em uma infraestrutura crítica real e brasileira permitirá contribuir no engajamento entre a academia, a indústria e o estado no esforço de soluções de proteção cibernética.

Apesar da plataforma permitir a detecção sistêmica, mais orientada para casos gerais, o HAI permite uma análise mais específica quanto às classes de ataque, seja quanto aos processos (P1 a P3) quanto ao tipo do alvo do ataque (variáveis de controle, de processo ou de configuração). Nesse caso, sugere-se aplicar técnicas orientadas à clusterização em conjunto à classificação.

Este trabalho não foi focado na interação entre a rede TI e a rede TO, porém conforme as simulações apresentem cenários que integrem ambas redes, espera-se que as detecções apresentem resultados ainda mais promissores.

Foi verificada uma oportunidade nos trabalhos relacionados ao HAI quanto à pesquisa de detecção *online* ou *stream detection* e *drift detection*, mais orientadas a uma aplicação de tempo real para cenários em que os ataques ensejem danos críticos imediatos.

Por fim, neste trabalho, a detecção em duas fases se deu pela composição de uma soma das predições supervisionada e não supervisionada, porém outras estratégias podem ser empregadas, envolvendo pesos e votações dos algoritmos, bem como avaliação por outras métricas menos usuais da literatura.

REFERÊNCIAS

AGYEPONG, E. *et al.* Towards a framework for measuring the performance of a security operations center analyst. *In*: **2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)**. Nova Iorque, EUA: IEEE, 2020.

AL-ABASSI, A. *et al.* An ensemble deep learning-based cyber-attack detection in industrial control system. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 8, p. 83965–83973, 2020. ISSN 2169-3536.

ALANAZI, M.; MAHMOOD, A.; CHOWDHURY, M. J. M. SCADA vulnerabilities and attacks: A review of the state-of-the-art and open issues. **Computers & Security**, Elsevier BV, v. 125, p. 103028, feb 2023.

ALMALAWI, A. *et al.* An unsupervised anomaly-based detection approach for integrity attacks on scada systems. **Computers & Security**, Elsevier BV, v. 46, p. 94–110, out. 2014. ISSN 0167-4048.

ANI, U. P. D.; HE, H. M.; TIWARI, A. Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective. **Journal of Cyber Security Technology**, Informa UK Limited, v. 1, n. 1, p. 32–74, nov 2016.

ANTROBUS, R. *et al.* Simaticscan: Towards a specialised vulnerability scanner for industrial control systems. *In*: **Electronic Workshops in Computing**. Swindon,United Kingdom: BCS Learning & Development, 2016. ISSN 1477-9358.

AZAMBUJA, A. J. G. d.; ALMEIDA, V. R. Um estudo bibliométrico das publicações sobre segurança cibernética na indústria 4.0. **Research, Society and Development**, Research, Society and Development, v. 10, n. 3, p. 4210312937e, mar. 2021. ISSN 2525-3409.

BAE, S.; HWANG, C.; LEE, T. Research on improvement of anomaly detection performance in industrial control systems. *In*: **Lecture Notes in Computer Science**. Berlin, Heidelberg: Springer International Publishing, 2021. p. 76–87. ISBN 9783030894320. ISSN 1611-3349.

BRASIL. Gabinete de Segurança Institucional da Presidência da República. **Portaria nº 2-GSI/PR, de 8 de fevereiro de 2008**, Imprensa Nacional, Brasília. DF, fev. 2008.

BRASIL. Ministério da Defesa. **Livro Branco da Defesa Nacional**, Brasília, DF, 2012.

BRASIL. Ministério da Defesa. **Estratégia Nacional de Defesa**, Imprensa Nacional, Brasília, DF, 2016.

BRASIL. Gabinete de Segurança Institucional da Presidência da República. **Decreto nº 9.573, de 22 de novembro DE 2018**, Imprensa Nacional, Brasília, DF, nov. 2018.

BUCZAK, A. L.; GUVEN, E. A survey of data mining and machine learning methods for cyber security intrusion detection. **IEEE Communications Surveys & Tutorials**, Institute of Electrical and Electronics Engineers (IEEE), Nova Iorque, EUA, v. 18, n. 2, p. 1153–1176, 2016.

CALDERS, T.; JAROSZEWICZ, S. Efficient auc optimization for classification. *In*: KOK, J. N. *et al.* (Ed.). **Knowledge Discovery in Databases: PKDD 2007**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 42–53. ISBN 9783540749769. ISSN 1611-3349.

CANADA. **Cyber Threat Bulletin: Cyber Threat Activity Related to the Russian Invasion of Ukraine**. Ottawa,Ontario, 2022. Disponível em: https://www.cyber.gc.ca/en/guidance/cyber-threat-bulletin-cyber-threat-activity-related-russian-invasion-ukraine. Acesso em: 21 jan. 2024.

CHAWLA, N. V. *et al.* Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, American Association for Artificial Intelligence, v. 16, p. 321 – 357, 2002.

CHEMINOD, M.; DURANTE, L.; VALENZANO, A. Review of security issues in industrial networks. **IEEE Transactions on Industrial Informatics**, Institute of Electrical and Electronics Engineers (IEEE), v. 9, n. 1, p. 277–293, fev. 2013. ISSN 1941-0050.

CHOI, S.; YUN, J.-H.; KIM, S.-K. A comparison of ics datasets for security research based on attack paths. *In*: LUIIJF, E.; ŽUTAUTAITĖ, I.; HÄMMERLI, B. M. (Ed.). **Lecture Notes in Computer Science**. Cham: Springer International Publishing, 2018. p. 154–166. ISBN 9783030058494. ISSN 1611-3349.

DEMERTZIS, K. *et al.* Variational restricted boltzmann machines to automated anomaly detection. **Neural Computing and Applications**, Springer Science and Business Media LLC, v. 34, n. 18, p. 15207–15220, mar. 2022. ISSN 1433-3058.

DZUNG, D. *et al.* Security for industrial communication systems. **Proceedings of the IEEE**, Institute of Electrical and Electronics Engineers (IEEE), v. 93, n. 6, p. 1152–1177, jun. 2005. ISSN 0018-9219.

ENGELEN, J. E. van; HOOS, H. H. A survey on semi-supervised learning. **Machine Learning**, Springer Science and Business Media LLC, v. 109, n. 2, p. 373–440, nov. 2019. ISSN 1573-0565.

EUA. **H.R.3359 - Cybersecurity and Infrastructure Security Agency Act of 2018**. 2018. 115th Congress Public Law 278. Disponível em: https://www.congress.gov/bill/115th-congress/house-bill/3359. Acesso em: 03 jan. 2024.

FRANÇA. **Loi n° 2018-607 du 13 juillet 2018**. Relative à la programmation militaire pour les années 2019 à 2025 et portant diverses dispositions intéressant la défense. Paris: Journal Officiel de la République Française, 2018.

FU, Y.; XUE, F. Mad: Self-supervised masked anomaly detection task for multivariate time series. *In*: **2022 International Joint Conference on Neural Networks (IJCNN)**. Nova Iorque, EUA: IEEE, 2022.

GOH, J. *et al.* A dataset to support research in the design of secure water treatment systems. *In*: **Lecture Notes in Computer Science**. Cham: Springer International Publishing, 2017. p. 88–99. ISBN 9783319713687. ISSN 1611-3349.

GRAHAM, J.; HIEB, J.; NABER, J. Improving cybersecurity for industrial control systems. *In*: **2016 IEEE 25th International Symposium on Industrial Electronics (ISIE)**. Nova Iorque, EUA: IEEE, 2016.

HAN, S.; WOO, S. S. Learning sparse latent graph representations for anomaly detection in multivariate time series. *In*: **Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2022. (KDD '22).

HOI, S. C. *et al.* Online learning: A comprehensive survey. **Neurocomputing**, Elsevier BV, v. 459, p. 249–289, out. 2021. ISSN 0925-2312.

HUDA, S. *et al.* Defending unknown attacks on cyber-physical systems by semi-supervised approach and available unlabeled data. **Information Sciences**, Elsevier BV, v. 379, p. 211–228, fev. 2017. ISSN 0020-0255.

ITAIPU; EXÉRCITO; PTI-BR, F. P. T. I. **Acordo de Parceria para Pesquisa, Desenvolvimento e Inovação nº ITAIPU 4500065528 e nº EME 22-DCT-003-00**. Diário Oficial da União, 2022. 26 p. Disponível em: https://www.in.gov.br/web/dou/-/extrato-de-parceria-399778022. Acesso em: 31 jan. 2024.

JENI, L. A.; COHN, J. F.; TORRE, F. D. L. Facing imbalanced data–recommendations for the use of performance metrics. *In*: **2013 Humaine Association Conference on Affective Computing and Intelligent Interaction**. Nova Iorque, EUA: IEEE, 2013.

JUBA, B.; LE, H. S. Precision-recall versus accuracy and the role of large data sets. **Proceedings of the AAAI Conference on Artificial Intelligence**, Association for the Advancement of Artificial Intelligence (AAAI), v. 33, n. 01, p. 4039–4048, jul. 2019. ISSN 2159-5399.

KABORE, R. *et al.* Review of anomaly detection systems in industrial control systems using deep feature learning approach. **Engineering**, Scientific Research Publishing, Inc., v. 13, n. 01, p. 30–44, 2021. ISSN 1947-394X.

KASPERSKY, I. C. **Threat landscape for industrial automation systems: Statistics for H1 2023**. Moscow, Russian Federation, 2023. Disponível em: https://ics-cert.kaspersky.com/publications/reports/2023/09/13/threat-landscape-for-industrial-automation-systems-statistics-for-h1-2023/. Acesso em: 21 nov. 2023.

KIM, D.; LEE, T. Anomaly detection in time-series data environment. *In*: **Proceedings of the 2020 ACM International Conference on Intelligent Computing and Its Emerging Applications**. New York, NY, USA: Association for Computing Machinery, 2021. (ACM ICEA '20). ISBN 9781450383042.

KIM, Y.; KIM, H. Cluster-based deep one-class classification model for anomaly detection. **Journal of Internet Technology**, Angle Publishing Co., Ltd., v. 22, n. 4, p. 903–911, jul. 2021. ISSN 1607-9264.

KIM, Y. G. *et al.* Revitalizing self-organizing map: Anomaly detection using forecasting error patterns. *In*: **IFIP Advances in Information and Communication Technology**. Cham: Springer International Publishing, 2021. p. 382–397. ISBN 9783030781200. ISSN 1868-422X.

KOAY, A. M. Y. *et al.* Machine learning in industrial control system (ics) security: current landscape, opportunities and challenges. **Journal of Intelligent Information Systems**, Springer Science and Business Media LLC, v. 60, n. 2, p. 377–405, out. 2022. ISSN 1573-7675.

KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. E. Data preprocessing for supervised leaning. **International journal of computer science**, v. 1, n. 2, p. 111–117, 2006.

KUMAR, A.; CHOI, B. J. Benchmarking machine learning based detection of cyber attacks for critical infrastructure. *In*: **2022 International Conference on Information Networking (ICOIN)**. Nova Iorque, EUA: IEEE, 2022.

LAI, K.-H. *et al.* Tods: An automated time series outlier detection system. **Proceedings of the AAAI conference on artificial intelligence**, v. 35, n. 18, p. 16060–16062, maio 2021.

LI, Z. *et al.* Copod: Copula-based outlier detection. *In*: **2020 IEEE International Conference on Data Mining (ICDM)**. Nova Iorque, EUA: IEEE, 2020.

LIANG, G. *et al.* A review of false data injection attacks against modern power systems. **IEEE Transactions on Smart Grid**, v. 8, n. 4, p. 1630–1638, 2017.

MAHDAVIFAR, S.; GHORBANI, A. A. Application of deep learning to cybersecurity: A survey. **Neurocomputing**, Elsevier BV, v. 347, p. 149–176, jun. 2019. ISSN 0925-2312.

MO, Y.; SINOPOLI, B. False data injection attacks in control systems. *In*: **Preprints of the 1st workshop on Secure Control Systems**. Estocolmo, Suécia: Association for Computing Machinery (ACM), 2010. v. 1.

MOKHTARI, S. *et al.* A machine learning approach for anomaly detection in industrial control systems based on measurement data. **Electronics**, MDPI AG, v. 10, n. 4, p. 407, feb 2021.

MUBARAK, S. *et al.* Anomaly detection in ics datasets with machine learning algorithms. **Computer Systems Science and Engineering**, Computers, Materials and Continua (Tech Science Press), v. 37, n. 1, p. 33–46, 2021. ISSN 0267-6192.

NAFEES, M. N. *et al.* Smart grid cyber-physical situational awareness of complex operational technology attacks: A review. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 55, n. 10, p. 1–36, fev. 2023. ISSN 1557-7341.

NICOLAIO, I. G. A.; MUNARETTO, A.; FONSECA, M. Uma proposta de detecção de ataques cibernéticos em sistemas de controle industrial (ics). *In*: **Anais do XXVIII Workshop de Gerência e Operação de Redes e Serviços (WGRS 2023)**. Porto Alegre: Sociedade Brasileira de Computação - SBC, 2023. (WGRS 2023).

PAN, S.; ADHIKARI, U. A specification-based intrusion detection framework for cyber-physical environment in electric power system. **Int. J. Netw. Secur.**, v. 17, n. 2, p. 174–188, 2015.

PANG, J.; PU, X.; LI, C. A hybrid algorithm incorporating vector quantization and one-class support vector machine for industrial anomaly detection. **IEEE Transactions on Industrial Informatics**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2022.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PLOKHY, S. **The Russo-Ukrainian War: From the bestselling author of Chernobyl**. Londres, Reino Unido: Penguin Books Limited, 2023. ISBN 9781802061796.

PURSIAINEN, C. Russia's critical infrastructure policy: What do we know about it? **European Journal for Security Research**, Springer Science and Business Media LLC, v. 6, n. 1, p. 21–38, out. 2020. ISSN 2365-1695.

RADOGLOU-GRAMMATIKIS, P. I.; SARIGIANNIDIS, P. G. Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 7, p. 46595–46620, 2019. ISSN 2169-3536.

REINO UNIDO. HM Government. **Integrated Review Refresh 2023: Responding to a more contested and volatile world**, Crown UK, Londres, Inglaterra, mar. 2023.

SAYGHE, A. *et al.* Survey of machine learning methods for detecting false data injection attacks in power systems. **IET Smart Grid**, v. 3, n. 5, p. 581–595, 2020.

SEONG, C. *et al.* Towards building intrusion detection systems for multivariate time-series data. *In*: **Silicon Valley Cybersecurity Conference**. Cham: Springer International Publishing, 2022. p. 45–56. ISBN 9783030960575. ISSN 1865-0937.

SHIN, H.-K. *et al.* **HAI - HIL-Based Augmented ICS Security Dataset**. 2023. Disponível em: https://github.com/icsdataset/hai. Acesso em: 12 dez. 2023.

SHIN, H.-K. *et al.* Hai 1.0: Hil-based augmented ics security dataset. *In*: **Proceedings of the 13th USENIX Conference on Cyber Security Experimentation and Test**. Berkeley, Califórnia: USENIX Association, 2020.

SPAFFORD, E.; METCALF, L.; DYKSTRA, J. **Cybersecurity Myths and Misconceptions Avoiding the Hazards and Pitfalls That Derail Us**: Avoiding the hazards and pitfalls that derail us. Boston: Pearson Education, 2023. ISBN 9780137929153.

STOUFFER, K. **Guide to Operational Technology (OT) Security**. Gaithersburg: National Institute of Standards and Technology, 2023.

SUDAR, K. *et al.* Analysis of cyberattacks and its detection mechanisms. *In*: **2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)**. Nova Iorque, EUA: IEEE, 2020.

SUNDARAMURTHY, S. C. *et al.* A human capital model for mitigating security analyst burnout. *In*: **Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)**. Ottawa: USENIX Association, 2015. p. 347–359. ISBN 978-1-931971-249.

TAX, D. **One-class classification; concept-learning in the absence of counter-examples**. 2001. Tese (Doutorado) — Delft University of Technology, 2001. ASCI Dissertation Series 65.

THOMAS, A. **Security Operations Center - Analyst Guide: SIEM Technology, Use Cases and Practices**. Scotts Valley,Califórnia: CreateSpace Independent Publishing Platform, 2016. ISBN 9781533408501.

VIEGAS, E. K.; SANTIN, A. O. Towards reliable intrusion detection in high speed networks. *In*: **Anais Estendidos do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg Estendido 2020)**. Porto Alegre: Sociedade Brasileira de Computação - SBC, 2020.

WANG, Y. *et al.* Monitoring industrial control systems via spatio-temporal graph neural networks. **Engineering Applications of Artificial Intelligence**, Elsevier BV, v. 122, p. 106144, jun. 2023. ISSN 0952-1976.

WILLIAMS, T. J. The purdue enterprise reference architecture. **Computers in Industry**, Elsevier BV, v. 24, n. 2–3, p. 141–158, set. 1994. ISSN 0166-3615.

XINGCHAO, B. Detecting anomalies in time-series data using unsupervised learning and analysis on infrequent signatures. **Journal of IKEEE**, Korean Institute of Electrical and Electronics Engineers, v. 24, n. 4, p. 1011–1016, 12 2020.

XUE, F.; YAN, W. Multivariate time series anomaly detection with few positive samples. *In*: **2022 International Joint Conference on Neural Networks (IJCNN)**. Nova Iorque, EUA: IEEE, 2022.

YAACOUB, J.-P. A. *et al.* Cyber-physical systems security: Limitations, issues and future trends. **Microprocessors and Microsystems**, Elsevier BV, v. 77, p. 103201, set. 2020. ISSN 0141-9331.

ZHAO, Y. *et al.* LSCP: Locally selective combination in parallel outlier ensembles. *In*: **Proceedings of the 2019 SIAM International Conference on Data Mining**. Filadélfia: Society for Industrial and Applied Mathematics, 2019. p. 585–593.