Análise do impacto de ataques adversários na detecção de intrusão em sistemas ciberfísicos

Antonio Carlos Campos da Silva Junior

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2024

Antonio Carlos Campos da Silva Junior

Análise do impacto de ataques adversários na detecção de intrusão em sistemas ciberfísicos

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Rodrigo Sanches Miani
Coorientador: Prof. Dr. Renan Gonçalves Cattelan

Uberlândia
2024

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

S586a
2024

Silva Junior, Antonio Carlos Campos da, 1996-
Análise do impacto de ataques adversários na detecção de intrusão em sistemas ciberfísicos [recurso eletrônico] / Antonio Carlos Campos da Silva Junior. - 2024.

Orientador: Rodrigo Sanches Miani.
Coorientador: Renan Gonçalves Cattelan.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Programa de Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2025.5044
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Miani, Rodrigo Sanches, 1983-, (Orient.). II. Cattelan, Renan Gonçalves, 1980-, (Coorient.). III. Universidade Federal de Uberlândia. Programa de Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

André Carlos Francisco
Bibliotecário-Documentalista - CRB-6/3408

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 49/2024, PPGCO				
Data:	20 de dezembro de 2024	Hora de início:	14:05	Hora de encerramento:	16:40
Matrícula do Discente:	12212CCP002				
Nome do Discente:	Antonio Carlos Campos da Silva Junior				
Título do Trabalho:	Análise do impacto de ataques adversários na detecção de intrusão em sistemas ciberfísicos				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Sistemas de Computação				
Projeto de Pesquisa de vinculação:	----------				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Renan Gonçalves Cattelan (Coorientador) FACOM/UFU, Elaine Ribeiro de Faria Paiva - FACOM/UFU, Juliano Fontoura Kazienko - UFSM e Rodrigo Sanches Miani - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Elaine Ribeiro de Faria Paiva - Birmingham/UK e Juliano Fontoura Kazienko -Santa Maria/RS . Os outros membros da banca e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Rodrigo Sanches Miani, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação

interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Rodrigo Sanches Miani**, **Professor(a) do Magistério Superior**, em 26/12/2024, às 10:56, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Renan Gonçalves Cattelan**, **Professor(a) do Magistério Superior**, em 26/12/2024, às 11:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Juliano Fontoura Kazienko**, **Usuário Externo**, em 26/12/2024, às 12:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Elaine Ribeiro de Faria Paiva**, **Professor(a) do Magistério Superior**, em 05/01/2025, às 19:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5948310** e o código CRC **DB2D0ADF**.

Referência: Processo nº 23117.085653/2024-24 SEI nº 5948310

Aos meus pais, vocês são a base de tudo.

Agradecimentos

Em primeiro lugar, agradeço a Deus, pois somente por Sua graça tudo isso foi possível. Agradeço também aos meus pais, que sempre me apoiaram e incentivaram a estudar, destacando a importância do conhecimento para o desenvolvimento pessoal e profissional, bem como para transformar vidas e impactar o meio em que vivemos. À minha namorada, Tayna Tamires, expresso minha gratidão por nunca me deixar desistir, por me motivar em cada dificuldade, por estar ao meu lado nos momentos difíceis e por celebrar comigo as conquistas. Agradeço ainda aos meus irmãos e à minha filha, por serem tão especiais e presentes em minha jornada. Não poderia deixar de mencionar meus amigos — os da escola, da faculdade e do trabalho —, com destaque para o grupo MMIV, que esteve comigo ao longo desta caminhada. Além disso, minha gratidão se estende a todos os meus familiares e colegas, que, de alguma forma, contribuíram para que este momento se tornasse realidade. Por fim, agradeço a todos os meus professores que me ajudaram e tiveram o máximo de paciência comigo em todo o processo.

"Ainda que eu ande pelo vale da sombra da morte, não temerei mal nenhum, porque tu estás comigo; o teu bordão e o teu cajado me consolam."
(Salmos 23:4)

Resumo

Sistemas ciberfísicos, do inglês *Cyber-Physical Systems* (CPSs), integram ecossistemas tecnológicos complexos, conectando computação, redes e processos físicos por meio de dispositivos interligados. Garantir a segurança da informação nesses sistemas é essencial, e algoritmos de aprendizado de máquina têm sido amplamente empregados para treinar Sistemas de Detecção de Intrusão (IDSs, do inglês *Intrusion Detection Systems*) baseados em anomalias, com o objetivo de proteger esses ecossistemas. Este estudo avalia o impacto de ataques adversários em algoritmos de aprendizado de máquina aplicados a IDSs baseados em anomalias, utilizando dois conjuntos de dados: *Power System Smart Grid Monitoring Power* e *Ereno IEC-61850 Intrusion Dataset*. A pesquisa aborda tanto a abordagem de classificador individual, do inglês *single classifier*, quanto a de comitê de classificadores, do inglês *ensemble classifier*, considerando dois tipos de ataques adversários: o *Fast Gradient Sign Method* (FGSM) e o *Jacobian-based Saliency Map Attack* (JSMA). Foram realizadas análises abrangentes, incluindo: comparação entre os ataques FGSM e JSMA, análise do impacto entre as classes de algoritmos de aprendizado de máquina (individual x comitê), efeitos da diminuição do tamanho do conjunto de treino roubado e impacto da inserção de amostras adversárias no conjunto de treino. Os resultados indicam que o impacto dos ataques adversários varia de acordo com o tipo de classificador e também com o conjunto de dados usado. Além disso, os algoritmos do grupo *ensemble classifier* demonstraram, de maneira geral, maior resistência aos ataques adversários. Observou-se também que a diminuição do tamanho do conjunto de treino roubado influenciou mais intensamente o impacto dos ataques FGSM em relação ao desempenho de linha de base dos classificadores. Em contrapartida, em determinados cenários, a inserção de amostras adversárias no conjunto de treino resultou em melhorias no desempenho dos classificadores. Em síntese, os resultados encontrados evidenciam que o ataque FGSM possui maior impacto negativo no desempenho dos IDSs. Ademais, os algoritmos do tipo comitê de classificadores se mostraram mais robustos contra ataques adversários quando comparados aos algoritmos do tipo classificador individual, consolidando sua eficácia em IDSs para CPSs.

Palavras-chave: Sistemas Ciber-Físicos (CPSs), Sistemas de Detecção de Intrusão (IDSs), Aprendizado de Máquina (AM), Classificador individual, Comitê de classificadores, Ataques Adversários, *Single Classifier, Ensemble Classifier, Fast Gradient Sign Method (FGSM) , Jacobian-based Saliency Map Attack (JSMA)* .

Abstract

Cyber-Physical Systems (CPSs) are complex technological ecosystems that integrate computing, networking, and physical processes through interconnected devices. Ensuring information security in these systems is critical, and machine learning algorithms are widely employed to train anomaly-based Intrusion Detection Systems (IDSs) for their protection. This study evaluates the impact of adversarial attacks on machine learning algorithms applied to anomaly-based IDSs using two datasets: *Power System Smart Grid Monitoring Power* and *Ereno IEC-61850 Intrusion Dataset*. The research investigates both *single classifier* and *ensemble classifier* approaches, focusing on two adversarial attacks: the *Fast Gradient Sign Method* (FGSM) and the *Jacobian-based Saliency Map Attack* (JSMA). Comprehensive analyses were conducted, including comparisons between FGSM and JSMA attacks, assessments of *single classifier* versus *ensemble classifier* performance, evaluations of the effect of reducing the stolen training set size, and the impact of incorporating adversarial samples into the training set. The findings reveal that the impact of adversarial attacks varies with the classifier type and dataset. Notably, *ensemble classifiers* generally exhibited greater resistance to adversarial attacks. A significant degradation in baseline performance was observed for FGSM attacks as the stolen training set size decreased. Conversely, in some scenarios, incorporating adversarial samples into the training set enhanced classifier performance. In summary, FGSM attacks were found to have a more pronounced negative impact on IDS performance. Additionally, *ensemble classifiers* demonstrated superior robustness to adversarial attacks compared to *single classifiers*, highlighting their effectiveness in IDSs for CPSs.

Keywords: Cyber-Physical Systems (CPSs), Intrusion Detection Systems (IDSs), Machine Learning, Adversarial Attacks, Single Classifier, Ensemble Classifier, Fast Gradient Sign Method (FGSM) , Jacobian-based Saliency Map Attack (JSMA).

Lista de ilustrações

Lista de tabelas

Lista de siglas

AD Árvore de Decisão

AdaBoost *Adaptive Boosting*

AM *Aprendizado de Máquina*

AI *Aprendizado Indutivo*

AE *Exemplos Adversários*

CW *Carlini & Wagner Attack*

CPS *Cyber-Physical Systems*

DeepFool *Deep Fool Attack*

DDoS *Distributed Denial of Service*

DL *Deep Learning*

DMZ *Demilitarized Zone*

DNN *Deep Neural Network*

DCS *Sistemas de Controle Distribuído*

FGSM *Fast Gradient Sign Method*

GBM *Gradient Boosting Machine*

GAN *Generative Adversarial Network*

GOOSE *Generic Object Oriented Substation Event*

HIDS *Host-based Intrusion Detection System*

IA *Inteligência Artificial*

IDS *Intrusion Detection System*

IoT *Internet of Things*

ICPS *Industrial Cyber-Physical Systems*

J48 Árvore de Decisão J48

JSMA *Jacobian-based Saliency Map Attack*

KDD'99 *Knowledge Discovery and Data Mining 1999 Dataset*

KNN *K-Nearest Neighbor*

LDA *Linear Discriminant Analysis*

ML *Machine Learning*

MLP *Multilayer Perceptron*

NIDS *Network-based Intrusion Detection System*

NSL-KDD *Network Security Lab Knowledge Discovery Dataset*

RF *Random Forest*

SCI *Sistemas de Controle Industrial*

SG *Smart Grid*

SCADA *Sistemas de Controle de Supervisão e Aquisição de Dados*

SVM *Support Vector Machine*

UNSW-NB15 *UNSW Network-Based Dataset 15*

UMF *Unidades de Medição Fasorial*

VPN *Virtual Private Network*

XGB *Extreme Gradient Boosting*

ZOO *Zeroth-Order Optimization*

Sumário

Capítulo 1

Introdução

Sistemas ciberfísicos, do inglês *Cyber-Physical Systems* (CPS), e Internet das Coisas, do inglês *Internet of Things* (IoT), são conceitos sobrepostos, ambos relacionados à integração digital de recursos. Essas definições abrangem a conectividade de rede e a capacidade computacional combinadas a dispositivos e sistemas físicos (GREER et al., 2019). Exemplos desses sistemas podem ser encontrados em setores diversos, como saúde, agricultura e energia. Nesse contexto, entre os tipos de CPS, destacam-se os *Sistemas de Controle Industrial*s (SCIs), que desempenham um papel importante na automação e monitoramento de processos industriais.

Segundo Anthi et al. (2021), os SCI desempenham uma função relevante na *Infraestrutura Nacional Crítica* (INC), atuando em setores como manufatura, estações de tratamento de água, refinarias de gás e petróleo, assistência médica e redes elétricas inteligentes, também conhecidas como *Smart Grid* (SG). Com o avanço tecnológico o mundo tornou-se cada vez mais interligado, surgindo a necessidade de expor os componentes dos SCI, conectando-os à Internet. Isso permite acesso remoto, funcionalidades de monitoramento, entre outros benefícios. No entanto, junto com o acesso à rede surgiram as vulnerabilidades relacionadas à segurança da informação nesse ambiente (KRAVCHIK; SHABTAI, 2018).

Proteger CPS contra ataques digitais tem se tornado uma preocupação crescente (MARTINS et al., 2020). Nesse contexto, destaca-se a importância dos Sistemas de Detecção de Intrusão, do inglês IDS, que representam a primeira linha de defesa ao proteger sistemas contra ataques cibernéticos (ALATWI; ALDWEESH, 2021).

De acordo com Sahani et al. (2023), IDSs baseados em *Aprendizado de Máquina* (AM), do inglês *Machine Learning* (ML), têm sido amplamente aplicados e evoluíram como mecanismos avançados de defesa em sistemas de última geração. Além disso, enquanto IDSs baseados em assinatura apresentam limitações na identificação de novos ataques, o uso de AM na construção de IDSs baseados em anomalias tem demonstrado potencial para superar essas restrições (ABUBAKAR; PRANGGONO, 2017). Dessa forma, IDSs têm a capacidade de processar informações de sistemas monitorados em tempo real, iden-

tificando indícios de comportamento anômalo. Com isso, podem gerar notificações aos responsáveis pelo sistema, permitindo uma tomada de decisão adequada. Esses fatores destacam as técnicas de AM como fundamentais no desenvolvimento de IDSs modernos (ZHANG et al., 2021).

De acordo com CORONA I.; GIACINTO (2013), os IDSs desempenham um papel essencial na camada de segurança da informação em diversos cenários computacionais, com o objetivo de impedir que esses sistemas sejam comprometidos por agentes maliciosos. No entanto, por integrarem a própria infraestrutura computacional, os IDSs também estão sujeitos a se tornarem alvos de ataques. Uma técnica frequentemente utilizada para burlar esses sistemas são os ataques adversários (CORONA I.; GIACINTO, 2013). No contexto de IDSs baseados em AM, o objetivo de um ataque adversário em IDSs é explorar vulnerabilidades do sistema, fazendo com que ataques reais sejam classificados como eventos legítimos ou benignos, permitindo que atividades maliciosas passem despercebidas. Isso é alcançado por meio da inserção de amostras adversárias, que consistem em entradas deliberadamente modificadas para induzir o erro nos IDSs (ALOTAIBI; RASSAM, 2023).

As ameaças enfrentadas pelos sistemas CPSs em setores críticos têm motivado estudos voltados ao desenvolvimento de diversos mecanismos de detecção de ataques, incluindo a detecção de anomalias baseada em modelos de AM. Embora a eficiência desses detectores seja avaliada com frequência por meio de conjuntos de testes com diferentes tipos de ataques, há uma limitação de cenários testados em relação aos ataques adversários (JIA et al., 2021). Apesar de ser um tema relativamente recente, ataques adversários em IDS aplicados a CPSs já estão sendo objeto de discussão.

Os trabalhos de Anthi et al. (2021), Alshahrani et al. (2022) e Silva, Miani e Zarpelao (2023) investigam o impacto dos ataques adversários em IDS aplicados a sistemas ciberfísicos. Anthi et al. (2021) aborda que a popularização de IDSs baseados em AM permitiu maior eficiência na detecção de ameaças em cenários de SCI. No entanto, a introdução dos IDSs também expôs esses sistemas a ataques adversários. Os autores de Anthi et al. (2021) utilizaram o projeto *CleverHans* [1] para gerar o ataque adversário JSMA e avaliar as mudanças nos classificadores *Random Forest* (RF) e Árvore de Decisão J48 (J48), utilizando a base de dados *Power System Smart Grid Monitoring Power*, desenvolvida pela *Mississippi State University* e *Oak Ridge National Laboratory*.

Em Silva, Miani e Zarpelao (2023), foram utilizados dois ataques adversários, JSMA e FGSM, em um CPS. Ambos os ataques foram implementados utilizando o projeto *CleverHans* e o framework *TensorFlow* para construir uma Perceptron Multi-Camadas, do inglês *Multilayer Perceptron* (MLP). Os classificadores testados foram RF, J48 e MLP, enquanto a base utilizada foi um conjunto de dados de energia gerado pela *Mississippi State University* e *Oak Ridge National Laboratory*.

Alguns aspectos relevantes não foram devidamente explorados nos trabalhos mencio-

[1] <https://github.com/cleverhans-lab/cleverhans>

nados, como a análise do desempenho de diferentes algoritmos de aprendizado de máquina ao considerar a discussão entre classificador individual, do inglês *single classifier*, e comitê de classificadores, do inglês *ensemble classifier*, expostos a amostras adversárias. Além disso, faltaram investigações mais aprofundadas sobre cenários envolvendo a variação dos parâmetros dos ataques adversários, a inserção de amostras adversárias no conjunto de treinamento e a avaliação desses fatores em diferentes bases de dados. Essas lacunas destacam a necessidade de estudos que ampliem a compreensão sobre o impacto desses ataques em diferentes abordagens e contextos.

1.1 Objetivos e desafios da pesquisa

O objetivo deste trabalho é avaliar o impacto de ataques adversários em IDSs aplicados a diferentes tipos de ambientes CPSs. Para atingir esse propósito, são analisados o desempenho de sete classificadores, divididos em dois grupos: classificador individual e comitê de classificadores. Adicionalmente, a pesquisa considera duas bases de dados e múltiplos cenários de ataques adversários.

Os objetivos específicos são:

- ❏ Estabelecer uma linha de base para avaliar o desempenho de IDSs baseados em anomalias em diferentes cenários de ataques adversários.
- ❏ Comparar o impacto dos ataques adversários JSMA e FGSM sobre o desempenho dos classificadores analisados.
- ❏ Avaliar o desempenho de classificadores individuais em relação a comitê de classificadores no contexto de detecção de ataques adversários.
- ❏ Analisar o impacto da redução da quantidade de amostras de treino roubadas pelo atacante no desempenho dos classificadores.
- ❏ Analisar os efeitos da inserção de amostras adversárias no conjunto de treino dos classificadores, considerando o impacto na detecção de ataques futuros.

1.2 Hipótese

A seguir são apresentadas as hipóteses estabelecidas para a execução do presente trabalho.

H1: Comparação entre os ataques JSMA e FGSM: Analogamente aos resultados obtidos em cenários diferentes de sistemas ciberfísicos (PACHECO; SUN, 2021), o impacto dos ataques adversários em IDS para tais ambientes varia de acordo com o conjunto de dados e também com o classificador utilizado.

H2: Comparação entre as classes de algoritmos (classificadores individuais/comitê de classificadores): Algoritmos baseados em comitê de classificadores possuem maior resistência contra ataques adversários em comparação aos algoritmos de aprendizado individual (classificadores individuais).

H3: Impacto da diminuição de amostras de treino roubadas: A redução na quantidade de amostras de treino roubadas diminui a vulnerabilidade do modelo de detecção de intrusão contra ataques adversários.

H4: Inserção de amostras adversárias no conjunto de treinamento: A inclusão de amostras adversárias no conjunto de treinamento contribui para a melhoria do desempenho do modelo contra novos ataques adversários de evasão, aumentando sua capacidade de generalização em cenários adversos.

1.3 Organização do trabalho

Esta dissertação está estruturada da seguinte forma. O Capítulo 2 apresenta a fundamentação teórica, oferecendo os conceitos e conhecimentos necessários para o desenvolvimento do trabalho. No Capítulo 3, discute os trabalhos relacionados que serviram como base bibliográfica e comparativa para esta pesquisa. As etapas do desenvolvimento deste trabalho são explicadas no Capítulo 4 que descreve detalhadamente as etapas e metodologias utilizadas no desenvolvimento deste trabalho. Os resultados obtidos são apresentados e discutidos no Capítulo 5 apresenta e analisa os resultados obtidos, discutindo suas implicações e relevância, e, por fim, o Capítulo 6 expõe as conclusões do estudo, ressaltando as contribuições, limitações e possíveis direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica necessária para compreender os conceitos e tecnologias que embasam esta pesquisa. A Seção 2.1, explora princípios e práticas voltados à proteção de sistemas e redes contra ameaças digitais. Em seguida, a Seção 2.2 descreve os mecanismos e abordagens utilizados para identificar atividades maliciosas em redes e dispositivos. Na Seção 2.3, são discutidos os métodos e algoritmos de AM aplicados à segurança, com foco em suas capacidades e limitações. A Seção 2.4 examina a integração entre componentes físicos e digitais, destacando os desafios de segurança específicos desses ambientes. Por fim, a Seção 2.5 analisa as ameaças que exploram vulnerabilidades dos sistemas ciberfísicos, abordando as técnicas e estratégias usadas para comprometer a segurança desses sistemas.

2.1 Segurança cibernética

Com a expansão da *Internet*, tornou-se comum que empresas e indivíduos armazenassem um grande volume de informações na rede. Esse aumento no fluxo de dados trouxe novos desafios relacionados à privacidade e à proteção dessas informações contra acessos indevidos. Nesse contexto, surgiu a necessidade de aprimorar a segurança da informação, destacando seu impacto significativo na proteção de dados.

Antes do advento da *Internet* e da popularização dos computadores pessoais, o armazenamento de informações era realizado principalmente por meio de papéis e documentos físicos. No entanto, a partir do século XXI, a informação passou a ser predominantemente armazenada em dispositivos computacionais. Dessa forma, o desenvolvimento de mecanismos para proteger essas informações digitais, denominado segurança da informação, consolidou-se como uma área de pesquisa de grande relevância (BENETTI, 2015). A segurança da informação baseia-se em três pilares fundamentais: confidencialidade, integridade e disponibilidade (CÔRTE, 2014).

1. Confidencialidade - Refere-se à garantia de que a informação será acessada apenas

por indivíduos devidamente autorizados, restringindo a visibilidade dos dados a pessoas ou entidades não autorizadas. Em outras palavras, busca assegurar que a informação permaneça sigilosa (BENETTI, 2015). A confidencialidade consiste em manter uma informação em segredo, independentemente de sua natureza. No entanto, sua manutenção pode ser onerosa, pois muitos dados são manipulados por pessoas vulneráveis a ataques de engenharia social e, frequentemente, estão inseridos em redes de computadores desprovidas de medidas adequadas de segurança e controle de acesso. A confidencialidade pode ser garantida por meio do uso de criptografia, ou seja, a aplicação de princípios e técnicas destinadas a tornar a informação ininteligível para aqueles que não possuem as chaves necessárias para decodificar o conteúdo (GUTTMAN; ROBACK, 1995).

2. Integridade - Relaciona-se à garantia de que o conteúdo da informação não será comprometido durante transações ou operações. Em outras palavras, assegura que os dados, mesmo que atualizados, removidos ou inseridos, permaneçam corretos e íntegros ao serem consultados. Mecanismos como algoritmos de *hash* e assinaturas digitais são amplamente utilizados para preservar a integridade dos dados, garantindo que não tenham sido alterados de forma não autorizada (BENETTI, 2015).

3. Disponibilidade - Refere-se à garantia de que a informação estará acessível sempre que requisitada, assegurando que esteja disponível para os usuários autorizados a qualquer momento. A disponibilidade depende de uma infraestrutura robusta, que envolve backups, redundância e proteção contra ataques de negação de serviço, do inglês *Distributed Denial of Service* (DDoS), visando garantir o acesso ininterrupto aos dados (BENETTI, 2015).

Além disso, outros dois conceitos relacionados à segurança da informação são citados por Guttman e Roback (1995): autenticidade, que se refere à capacidade de verificar quem enviou e quem recebeu uma mensagem, evitando, assim, o não repúdio, situação em que o autor não pode negar ter criado e assinado o documento, e responsabilização, que assegura que as ações de qualquer elemento dentro de um domínio possam ser atribuídas ao seu autor. A combinação desses cinco pilares forma a base para uma abordagem abrangente e eficiente de segurança da informação, protegendo os dados contra ameaças externas e internas de maneira integral. A Figura 1 ilustra os componentes da segurança da informação.

2.2 Sistemas de detecção de intrusão

A expansão da comunicação no século XXI criou diversas possibilidades de interação entre pessoas, independentemente da distância, por meio de redes de comunicação. Com

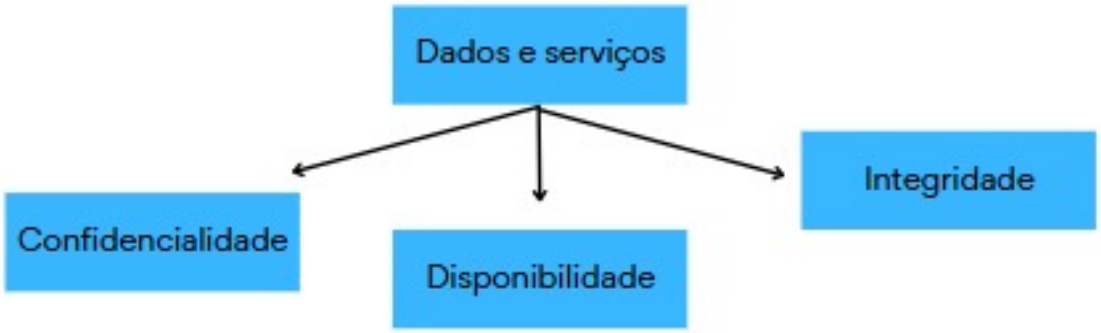

Figura 1 – Componentes da segurança da informação. Fonte: elaborado pelo autor (2024)

isso, questões como a segurança dos dados e das informações compartilhadas nessas interações têm recebido grande atenção dos pesquisadores. Um método de destaque para esse propósito é o Sistema de Detecção de Intrusão. IDSs são capazes de detectar intrusões na rede, auxiliando a reduzir ataques cibernéticos (EINY; OZ; NAVAEI, 2021).

Ao considerar IDSs, existem os sistemas baseados em *Host-based Intrusion Detection System* (HIDS), que monitoram as atividades de um único dispositivo ou servidor (*host*). Também existem os sistemas baseados em *Network-based Intrusion Detection System* (NIDS), que monitoram o tráfego de rede em tempo real para identificar atividades maliciosas. Além desses, há os sistemas híbridos, que integram os conceitos de HIDS e NIDS em uma única estrutura (KUMAR; SHARMA, 2018).

Nakamura e Geus (2007) discutem as diferentes posições que um IDS pode ocupar dentro de uma arquitetura de segurança de redes, como ilustrado na Figura 2:

- **IDS 1:** Este IDS identifica todas as tentativas de ataque direcionadas à rede da organização, incluindo aquelas que não causariam impacto, como tentativas de ataque a servidores Web inexistentes. Essa posição do IDS fornece uma valiosa fonte de dados sobre os tipos de ataques que poderiam ser direcionados à organização.

- **IDS 2:** Operando no próprio *firewall*, este IDS é capaz de detectar tentativas de ataque direcionadas especificamente ao *firewall*.

- **IDS 3:** Localizado para monitorar os servidores na Zona Desmilitarizada (*Demilitarized Zone* (DMZ)), este IDS detecta tentativas de ataque que conseguiram atravessar o *firewall*. Dessa forma, ele é capaz de identificar ataques contra serviços legítimos situados na DMZ.

- **IDS 4:** Este IDS detecta tentativas de ataque contra recursos internos que passaram pelo *firewall*, incluindo ataques que podem ocorrer por meio de uma Rede Privada Virtual (*Virtual Private Network* (VPN)).

- **IDS 5:** Responsável por monitorar os servidores na segunda Zona Desmilitarizada (DMZ 2), este IDS identifica tentativas de ataque que já passaram pelo *firewall*, pela VPN ou por algum serviço da DMZ 1, como o servidor Web. Isso ocorre porque os recursos na DMZ 2 não são acessíveis diretamente ao usuário, a menos que o acesso seja realizado via algum servidor na DMZ 1 ou por meio de VPN.

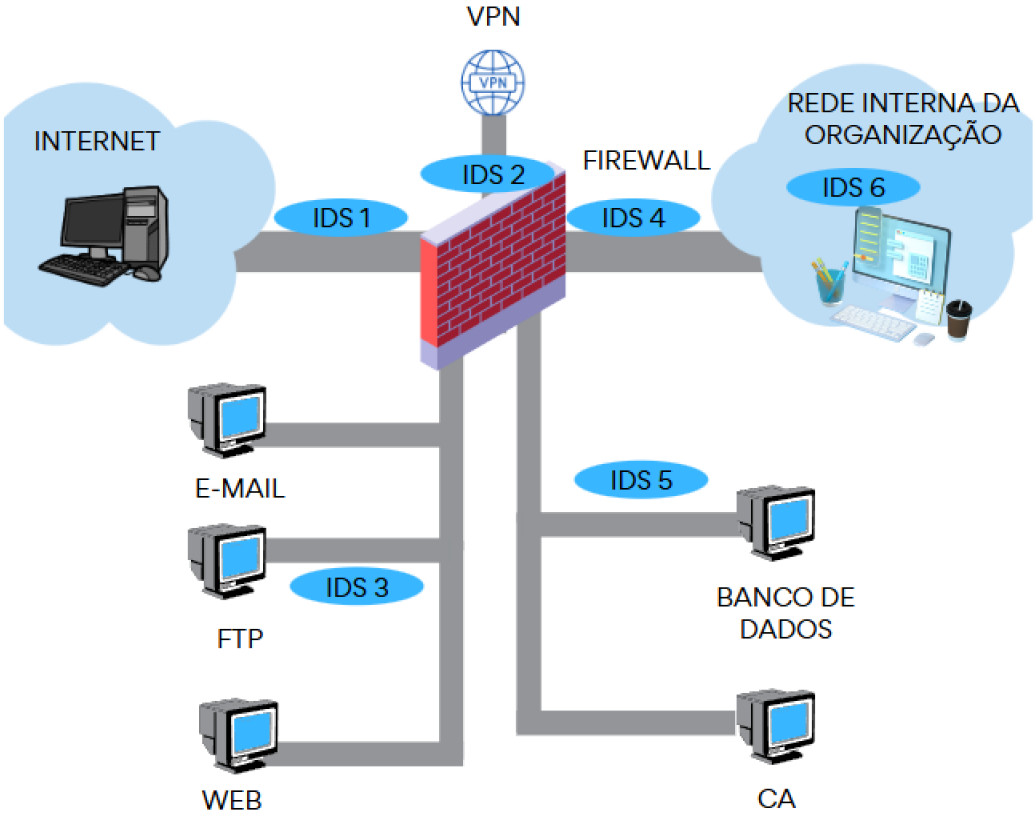

Figura 2 – Posicionamento de IDSs na rede. Fonte: Adaptado de (NAKAMURA; GEUS, 2007).

De acordo com Garcia-Teodoro et al. (2009), IDSs são ferramentas de segurança que buscam reforçar os mecanismos de proteção da informação. Essas ferramentas utilizam duas metodologias principais para a detecção de ataques (UTIMURA, 2020).

A primeira metodologia é a detecção por assinatura, na qual o IDS busca padrões de possíveis ações maliciosas em uma base de dados, chamados de assinaturas. Essa abordagem apresenta uma boa taxa de assertividade para ataques conhecidos; no entanto, é ineficaz contra ataques que não possuem registros na base de dados. Um dos principais

desafios desse modelo é manter a base de dados constantemente atualizada (UTIMURA, 2020).

A segunda metodologia é a detecção por anomalia. IDSs que são expostos a um processo de treinamento para distinguir tráfego de rede normal de tráfego malicioso são conhecidos como IDSs de detecção por anomalia. O principal desafio dessa abordagem é alcançar um alto nível de assertividade enquanto mantém um baixo índice de falsos positivos. Por não estarem limitados às assinaturas de ataques conhecidos, essa metodologia também é capaz de identificar ataques desconhecidos (NASCIMENTO; CORREIA, 2011).

No entanto, Rocha et al. (2023) apontam que identificar ataques que não foram contemplados durante a fase de treinamento pode ser um desafio significativo para IDSs baseados em algoritmos de AM. O estudo conclui que, na maioria dos casos, modelos supervisionados de detecção de intrusão não conseguem identificar ataques desconhecidos.

Karatas, Demir e Sahingoz (2020) destaca que, para proteger servidores e sistemas contra hackers que utilizam ataques desconhecidos, é possível empregar IDSs treinados com técnicas de AM, como, por exemplo, *K-Nearest Neighbor* (KNN), RF, *Gradient Boosting Machine* (GBM), *Adaptive Boosting* (AdaBoost), Árvore de Decisão (AD) e *Linear Discriminant Analysis* (LDA). Uma abordagem amplamente discutida na literatura para o desenvolvimento de IDSs baseados em anomalias é justamente a partir do uso de AM. Mais detalhes sobre esse aspecto serão apresentados na Seção 2.3.

2.3 Aprendizado de máquina

A partir do século XXI, surgiram conjuntos de dados caracterizados por um volume significativo de informações. Devido à grande quantidade de dados nesses conjuntos, eles passaram a ser denominados *Big Data*. Dada a complexidade no processamento e na manipulação desses dados, técnicas associadas ao conceito de *Inteligência Artificial* (IA) foram desenvolvidas para auxiliar no gerenciamento dessas informações, destacando-se o AM como a principal abordagem nesse contexto.

De acordo com Helm et al. (2020), o AM tem a capacidade de aprender e melhorar suas decisões por meio de algoritmos computacionais conhecidos como classificadores. Esses algoritmos utilizam conjuntos de dados de entrada e saída para "aprender" com os padrões analisados, adquirindo a capacidade de tomar decisões ou fazer recomendações de forma autônoma. Ainda segundo Helm et al. (2020), após processos que envolvem interações e ajustes nos classificadores geralmente denominados treinamento, a máquina desenvolve a habilidade de processar entradas e prever saídas (predição). A teoria do AM fundamenta-se nos princípios do *Aprendizado Indutivo* (AI), que pode ser dividido em três categorias: supervisionado, não supervisionado e semi-supervisionado (BRUNIALTI et al., 2015).

No contexto do Aprendizado Supervisionado, cada exemplo fornecido ao algoritmo é acompanhado de um rótulo correspondente, indicando a classe à qual o exemplo pertence. Um exemplo comum é o problema de classificação de imagens, em que se busca diferenciar entre imagens de gatos e cachorros. Esses exemplos são representados por vetores compostos por atributos e pelo rótulo associado à classe. O objetivo do algoritmo é desenvolver um classificador capaz de identificar corretamente a classe de novos exemplos ainda não rotulados. Quando os rótulos pertencem a categorias discretas, o problema é denominado classificação; já quando os valores são contínuos, trata-se de regressão (SILVA, 2023).

No contexto do Aprendizado Não Supervisionado, os exemplos são apresentados ao algoritmo sem a atribuição prévia de rótulos. Nesse método, o algoritmo busca identificar padrões nos atributos dos exemplos fornecidos, agrupando-os com base em suas similaridades. O objetivo principal é revelar estruturas subjacentes nos dados, organizando-os em agrupamentos ou *clusters*. Após a formação desses agrupamentos, é geralmente necessária uma análise complementar para interpretar e atribuir significados aos *clusters* no contexto do problema em estudo (SILVA, 2023). Complementarmente, a categoria de Aprendizado Semi-Supervisionado está relacionada à utilização de algoritmos híbridos que combinam a exploração de dados não rotulados com a aplicação de correções baseadas em erros e a maximização de medidas de qualidade, conforme descrito por (BRUNIALTI et al., 2015).

Além disso, os algoritmos podem ser classificados em dois grupos principais: o grupo de classificadores individuais, que consiste em modelos que utilizam apenas um classificador para realizar suas predições, e o grupo de comitê de classificadores, que emprega múltiplos classificadores para combinar suas predições. Um dos motivos para se estudar os algoritmos do grupo de comitê de classificadores é a possibilidade de que os classificadores individuais não apresentem a mesma eficácia quando comparados aos comitê de classificadores no contexto de uso em IDSs (HARIGUNA; HANANTO, 2022). Exemplos de algoritmos do grupo de classificadores individuais incluem *Support Vector Machine* (SVM), KNN e AD. Por outro lado, entre os algoritmos do grupo de comitê de classificadores, destacam-se RF, GBM, AdaBoost e *Extreme Gradient Boosting* (XGB).

Considerando o processo de AM, existem quatro etapas fundamentais: preprocessamento, treinamento, teste e avaliação. O preprocessamento tem como finalidade melhorar a qualidade dos dados (BATISTA, 2003). O treinamento é essencial para que um modelo de AM se torne o mais preciso possível. Esse processo se inicia com a divisão da base de dados em, no mínimo, duas partes: uma para treinamento e outra para teste.

De acordo com Uçar et al. (2020), o modelo é treinado considerando apenas os dados de treinamento. Após o treinamento, o modelo é testado com os dados de teste, que devem ser inéditos para o modelo. Assim, torna-se possível avaliar seu desempenho por meio dos critérios selecionados de acordo com o problema abordado. A avaliação de modelos de aprendizado de máquina consiste em medir o desempenho do modelo após o treinamento, utilizando um conjunto de dados de teste que não foi exposto durante o

processo de treinamento. O desempenho é analisado com base em critérios previamente estabelecidos. Essas etapas são ilustradas na Figura 3.

Figura 3 – Etapas do aprendizado de máquina. Fonte: elaborado pelo autor (2024).

Em AYUB Md Ahsan; JOHNSON (2020), evidencia-se que o uso de algoritmos de AM em IDSs é amplamente difundido e tem demonstrado resultados satisfatórios, contribuindo para a robustez e eficácia desses sistemas. Entretanto, Mohanty, Roudsari e Lashkari (2022) aponta que, embora os resultados indiquem certa eficácia na classificação do tráfego de rede (benigno/maligno), essa abordagem apresenta vulnerabilidades frente a ataques adversários.

2.4 Sistemas ciberfísicos

De acordo com Prandel (2022), um CPS, é caracterizado pela interação entre processos físicos e computacionais. O avanço tecnológico tem possibilitado a integração desses processos em diversas áreas, como saúde, educação, setor financeiro, gerenciamento de energia elétrica e recursos hídricos (CÂMARA, 2017). A Figura 4 apresenta exemplos de CPSs aplicados nos contextos de saúde, energia elétrica e educação.

Levando em consideração esse contexto, Prandel (2022) aponta que o relacionamento entre o físico e o digital pode ser considerado, atualmente, um dos pilares da indústria moderna. Além disso, os sistemas mencionados conseguem atuar em ambientes multi-espaciais e multi-temporais.

Nesse cenário, surgem conceitos importantes relacionados às bases de dados utilizadas. O primeiro deles é o já mencionado SG, que se refere a uma solução voltada para as redes elétricas. O SG propicia a integração da arquitetura entre os componentes envolvidos no sistema (LOPES et al., 2012). Ademais, essa arquitetura se baseia amplamente em redes de telecomunicações, trazendo vantagens como maior eficiência e confiabilidade para o sistema, além de possibilitar a comunicação entre dispositivos inteligentes na rede (LOPES et al., 2012).

Além disso, é fundamental considerar a norma IEC 61850, que estabelece um padrão capaz de equilibrar a comunicação entre os dispositivos do SG. Ainda segundo (LOPES

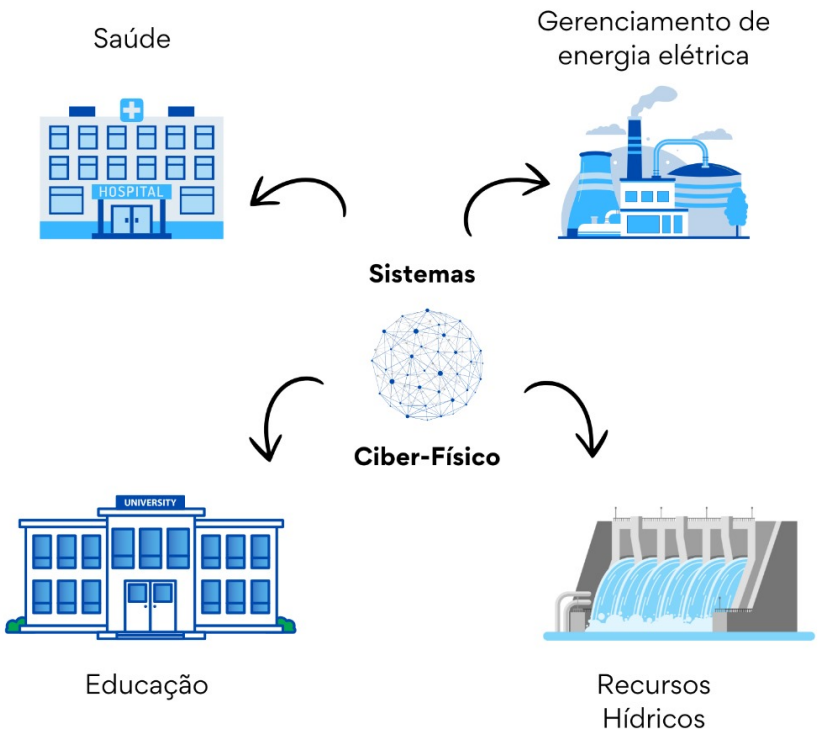

Figura 4 – Exemplo de sistemas ciberfísicos. Fonte: elaborado pelo autor (2024).

et al., 2012), é necessário abordar os distintos desafios destacados para o desenvolvimento de uma rede de geração para as SGs. Entre esses desafios, incluem-se a segurança, o gerenciamento, a qualidade do serviço e a migração tecnológica.

Em relação aos conceitos que se conectam com os sistemas ciberfísicos, o termo SCI refere-se a diferentes tipos de sistemas de controle, como os sistemas de controle de supervisão e aquisição de dados *Sistemas de Controle de Supervisão e Aquisição de Dados* (SCADA), sistemas de controle distribuído *Sistemas de Controle Distribuído* (DCS) e outras configurações, como controladores lógicos programáveis, amplamente utilizados em setores industriais e infraestruturas críticas. Um SCI consiste na combinação de componentes de controle (elétrica, mecânica, hidráulica, pneumática, por exemplo) que atuam em conjunto para alcançar um objetivo industrial, como a fabricação ou o transporte de matéria ou energia, de acordo com (STOUFFER et al., 2011).

De maneira geral, os CPSs representam um conceito amplo que abrange todas as aplicações que integram o físico e o digital. Dentro desse escopo, as SGs são uma aplicação

voltada para o setor de energia. As SGs utilizam SCIs para controlar os fluxos de energia, e a norma IEC 61850 define protocolos empregados nesses SCIs para padronizar a comunicação entre dispositivos. A Figura 5 ilustra esse cenário.

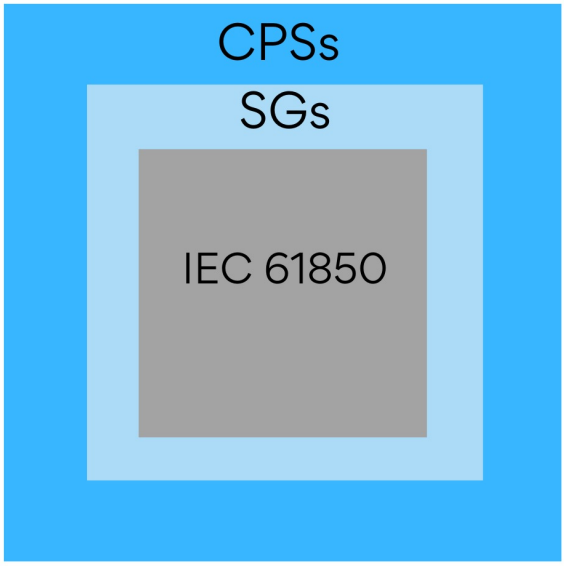

Figura 5 – Hierarquia dos sistemas ciberfísicos. Fonte: elaborado pelo autor (2024).

Os CPSs avançam em diversas áreas de aplicação críticas de segurança, sendo crucial garantir que eles operem sem causar danos às pessoas, ao meio ambiente e aos ativos. Contudo, a complexidade dos CPSs os torna vulneráveis e propensos a ataques cibernéticos (BOLBOT VICTOR; THEOTOKATOS, 2019). Segundo Alguliyev, Imamverdiyev e Sukhostat (2018), assegurar os fatores de segurança da informação em CPSs é um dos problemas mais complexos, devido à variedade de itens que fazem parte desse contexto.

Nesse cenário, algoritmos de AM estão sendo utilizados para a análise de segurança de sistemas ciberfísicos, tornando-se uma ferramenta essencial para a análise automática desses sistemas, permitindo o monitoramento contínuo e o aprimoramento das medidas de segurança (JAMAL et al., 2023). Rouzbahani et al. (2020) apresenta um estudo de caso que utiliza algoritmos de AM com o objetivo de detectar anomalias, demonstrando a eficácia dessas técnicas na classificação de dados maliciosos.

2.5 Ataques adversários

Apesar do sucesso dos modelos de AM na classificação de tráfego de rede, uma consequência típica do uso dessa abordagem é a sua fragilidade em relação a ataques adversários (HUANG et al., 2011). Segundo Martins et al. (2020), ataques adversários são

definidos como mecanismos que tentam burlar algoritmos de classificação, utilizando entradas projetadas especificamente para confundir os algoritmos de AM, resultando em predições incorretas. Para Alhajjar, Maxwell e Bastian (2021), os ataques adversários são eficazes em diversos cenários, como reconhecimento de imagem, reconhecimento de fala e detecção de spam.

Mohanty, Roudsari e Lashkari (2022) identificam três tipos principais de ataques adversários: envenenamento, evasão e roubo de modelo. O ataque de envenenamento ocorre quando alterações são realizadas nas amostras de treinamento, inserindo dados manipulados para comprometer o aprendizado do modelo. O ataque de evasão, por sua vez, manipula amostras de teste para enganar o modelo e induzi-lo a gerar saídas incorretas. Por fim, o ataque de roubo de modelo ocorre quando o atacante obtém informações privilegiadas sobre o modelo, visando replicá-lo ou explorá-lo. Esse tipo de ataque permite criar uma cópia funcional do modelo.

Existem três cenários principais para ataques adversários: *White Box*, *Gray Box* e *Black Box*. No cenário *White Box*, o atacante possui conhecimento detalhado sobre o modelo de classificação, incluindo o algoritmo de AM utilizado, a base de dados, e parte ou todos os parâmetros do modelo. Já no *Gray Box*, o invasor tem apenas conhecimento parcial ou limitado sobre a arquitetura do alvo. Por fim, caso o atacante não possua qualquer conhecimento sobre o modelo, o ataque é classificado como *Black Box* (AYUB MD AHSAN; JOHNSON, 2020). A Figura 6 ilustra cenários de ataque adversário aplicados a um IDS, destacando o tipo evasão. Dependendo do nível de conhecimento do atacante sobre o alvo, o ataque pode ser classificado como *White Box*, *Gray Box* ou *Black Box*. Complementando essa análise, a Tabela 1 sintetiza e exemplifica as diferenças fundamentais entre os cenários mencionados.

Um atacante busca introduzir amostras adversárias, que correspondem a uma entrada propositalmente modificada. Essas amostras são projetadas para alterar o comportamento do classificador, de forma que, dado um conjunto de dados com atributos x e rótulo y, o modelo passe a associar um rótulo incorreto à entrada modificada x', ainda que x' seja praticamente indiscriminável de x. Além disso, existem diversas técnicas para gerar essas amostras (ALOTAIBI; RASSAM, 2023)

Dentre as diversas técnicas de ataque adversário, destacam-se duas: o FGSM (GOODFELLOW; SHLENS; SZEGEDY, 2015) e o JSMA (PAPERNOT et al., 2016). Ambos os métodos utilizam a estratégia de adicionar pequenas perturbações à amostra original. Além disso, esses ataques geralmente fazem uso de uma MLP, um tipo de rede neural artificial composta por várias camadas de nós (neurônios) conectados entre si, que é previamente treinada, com os dados roubados do treinamento, como modelo no processo de geração das amostras adversárias.

O método FGSM tem como objetivo introduzir pequenos ruídos nos dados de entrada para que o classificador produza uma classificação diferente da esperada. O método

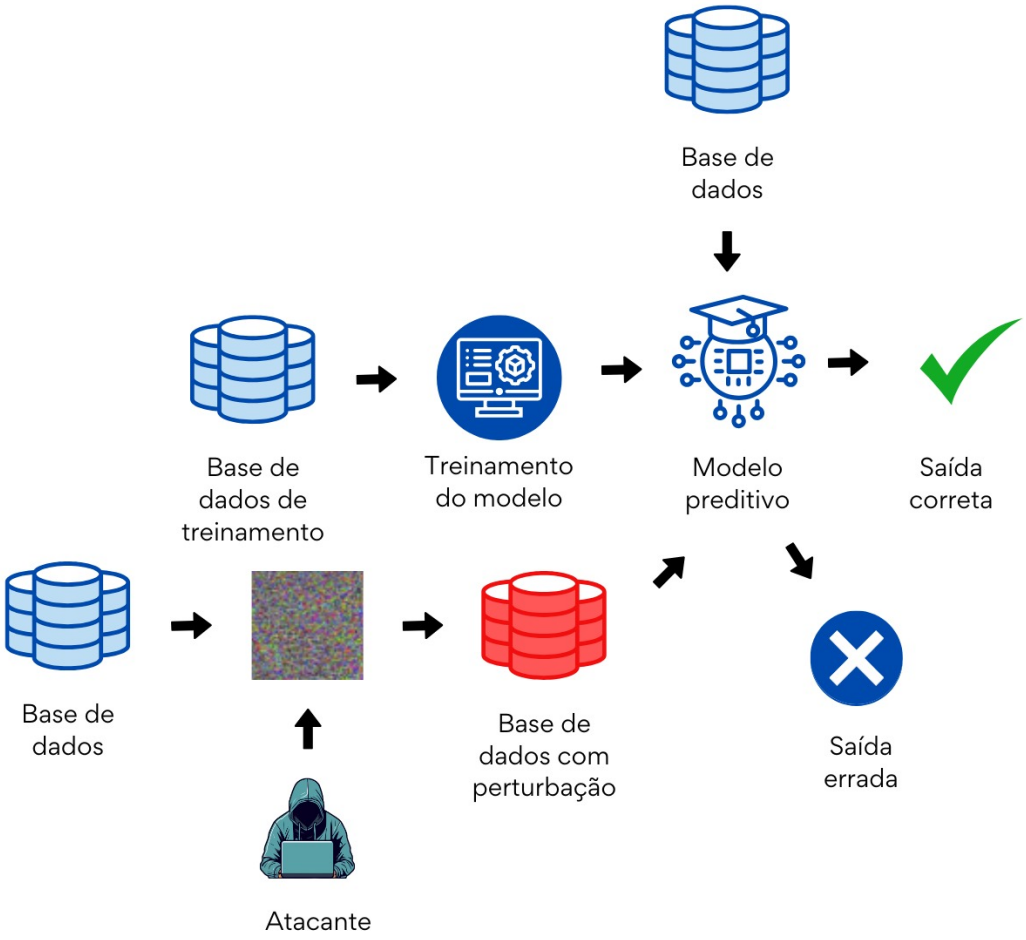

Figura 6 – *Pipeline* de um ataque adversário aplicado a um IDS, destacando o tipo evasão. Fonte: elaborado pelo autor (2024).

JSMA, por sua vez, também insere ruídos nos dados de entrada para induzir o erro no classificador, porém, de maneira mais direcionada. Com base em um mapa de saliência, esse ataque identifica as características que mais influenciam na tomada de decisão do modelo (PACHECO; SUN, 2021).

A Eq. 1 representa o ataque FGSM. $\mathbf{x}$ indica a entrada original; $\mathbf{x}^*$ caracteriza o exemplo adversário usado; ϵ é o parâmetro que indica o nível de perturbação adicionado à entrada original; $\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, \mathbf{y})$ indica o gradiente da função de perda J em relação à entrada $\mathbf{x}$; $J(\theta, \mathbf{x}, \mathbf{y})$: representa a função de perda; θ denota os parâmetros do modelo de aprendizado de máquina, como pesos e vieses em uma rede neural; $\mathbf{y}$ representa o rotulo correto associado a entrada verdadeira $\mathbf{x}$; $\cdot$sign representa a função sinal, que retorna o sinal de cada componente do gradiente.

$$\mathbf{x}^* = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, \mathbf{y})) \tag{1}$$

A Eq. 2 representa o ataque JSMA onde $\mathbf{x}$ indica a entrada original; $\mathbf{x}^*$ representa a

Tabela 1 – Diferenças entre cenários de ataque adversário *Gray Box*, *White Box* e *Black Box*.

Cenário	**Conhecimento do Atacante**	**Exemplo de Ataque**
White Box	Conhecimento completo do modelo: algoritmo, base de dados, parâmetros, etc.	Um funcionário da empresa tem acesso ao código-fonte, hiperparâmetros e dados de treinamento do modelo de detecção de intrusão. Com esse conhecimento, ele ajusta os parâmetros de FGSM para maximizar o impacto e comprometer o sistema.
Gray Box	Conhecimento parcial do modelo: apenas informações limitadas sobre a arquitetura ou funcionamento.	O atacante sabe que o modelo é baseado em AM, mas não conhece os pesos específicos ou a base de dados. Ele usa ataques genéricos como JSMA para enganar o sistema sem precisar de acesso direto ao modelo.
Black Box	Nenhum conhecimento do modelo: o atacante só observa as saídas para as entradas dadas.	O atacante treina um modelo substituto utilizando um conjunto de dados similar e gera exemplos adversários para enganar o modelo real, aproveitando a transferabilidade de ataques adversários de evasão.

entrada modificada; ϵ caracteriza o fator de escala; $\cdot$sign denota a função sinal; $\left(\frac{\partial F_t(\mathbf{x})}{\partial \mathbf{x}_i}\right)$ é o gradiente da função de saída.

$$\mathbf{x}^* = \mathbf{x} + \alpha \cdot \text{sign}\left(\frac{\partial F_t(\mathbf{x})}{\partial \mathbf{x}_i}\right) \tag{2}$$

Sobre a implementação dos ataques, a *Adversarial Robustness Toolbox* (ART) [1] é uma biblioteca desenvolvida pela IBM Research como software livre, utilizando Python. O ART está relacionado a muitos ataques adversários de última geração e a mecanismos de defesa para modelos convencionais de aprendizado de máquina e aprendizado profundo (WOLDEYOHANNES, 2021). A biblioteca pode ser utilizada como uma ferramenta de desenvolvimento e treinamento de modelos de aprendizado de máquina contra ataques adversários já citados, como evasão, envenenamento, extração e ataques de inferência. Além disso, pode ser usada para defender e mensurar a robustez dos modelos. Erdemir et

[1] https://github.com/Trusted-AI/adversarial-robustness-toolbox

al. (2021) destaca que alguns exemplos de metodologias usadas para a geração de ataques adversários utilizam os parâmetros padrão para os geradores de *Exemplos Adversários* (AE) dos métodos *Carlini & Wagner Attack* (CW), JSMA e *Deep Fool Attack* (DeepFool). Em 1, apresenta-se um exemplo de implementação do ataque JSMA. Além disso, em 2, apresenta-se um exemplo de implementação do ataque FGSM.

Listing 1 Exemplo de implementação do ataque JSMA em Python usando o pacote ART.

```python
# Importar bibliotecas necessárias
import numpy as np
from art.attacks.evasion import SaliencyMapMethod
from art.estimators.classification import KerasClassifier
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Flatten
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.utils import to_categorical

# Carregar e pré-processar os dados (ex.: MNIST)
dados_treino, rótulos_treino, dados_teste, rótulos_teste = carregar_dados_mnist()
rótulos_treino = to_categorical(rótulos_treino, num_classes=10)
rótulos_teste = to_categorical(rótulos_teste, num_classes=10)

# Normalizar os dados para o intervalo [0, 1]
dados_treino = dados_treino.astype("float32") / 255.0
dados_teste = dados_teste.astype("float32") / 255.0

# Configurar o modelo MLP
modelo_mlp = Sequential([
    Flatten(input_shape=(28, 28)),
    Dense(128, activation='relu'),
    Dense(64, activation='relu'),
    Dense(10, activation='softmax')
])

# Compilar o modelo
modelo_mlp.compile(optimizer=Adam(learning_rate=0.001),
                   loss='categorical_crossentropy',
                   metrics=['accuracy'])

# Treinar o modelo
modelo_mlp.fit(dados_treino, rótulos_treino, epochs=10, batch_size=64, verbose=1)

# Adaptar o modelo para ART
modelo_art = KerasClassifier(model=modelo_mlp, clip_values=(0, 1))

# Configurar o ataque JSMA
ataque_jsma = SaliencyMapMethod(
    classifier=modelo_art,
    theta=0.1,
    gamma=0.5,
    batch_size=1
)

# Gerar amostras adversárias no conjunto de teste
dados_adversarios = ataque_jsma.generate(x=dados_teste)

# Avaliar o desempenho do modelo
print("Desempenho no conjunto original:")
modelo_art.evaluate(dados_teste, rótulos_teste)

print("Desempenho no conjunto adversário:")
modelo_art.evaluate(dados_adversarios, rótulos_teste)
```

Listing 2 Exemplo de implementação do ataque FGSM em Python usando o pacote ART.

```
# Importar bibliotecas necessárias
import numpy as np
from art.attacks.evasion import FastGradientMethod
from art.estimators.classification import KerasClassifier
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Flatten
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.utils import to_categorical

# Carregar e pré-processar os dados (ex.: MNIST)
dados_treino, rótulos_treino, dados_teste, rótulos_teste = carregar_dados_mnist()
rótulos_treino = to_categorical(rótulos_treino, num_classes=10)
rótulos_teste = to_categorical(rótulos_teste, num_classes=10)

# Normalizar os dados para o intervalo [0, 1]
dados_treino = dados_treino.astype("float32") / 255.0
dados_teste = dados_teste.astype("float32") / 255.0

# Configurar o modelo MLP
modelo_mlp = Sequential([
    Flatten(input_shape=(28, 28)),
    Dense(128, activation='relu'),
    Dense(64, activation='relu'),
    Dense(10, activation='softmax')
])

# Compilar o modelo
modelo_mlp.compile(optimizer=Adam(learning_rate=0.001),
                   loss='categorical_crossentropy',
                   metrics=['accuracy'])

# Treinar o modelo
modelo_mlp.fit(dados_treino, rótulos_treino, epochs=10, batch_size=64, verbose=1)

# Adaptar o modelo para ART
modelo_art = KerasClassifier(model=modelo_mlp, clip_values=(0, 1))

# Configurar o ataque FGSM
ataque_fgsm = FastGradientMethod(
    classifier=modelo_art,
    eps=0.1  # Intensidade da perturbação adversária
)

# Gerar amostras adversárias no conjunto de teste
dados_adversarios = ataque_fgsm.generate(x=dados_teste)

# Avaliar o desempenho do modelo
print("Desempenho no conjunto original:")
modelo_art.evaluate(dados_teste, rótulos_teste)

print("Desempenho no conjunto adversário:")
modelo_art.evaluate(dados_adversarios, rótulos_teste)


```

Capítulo 3

Trabalhos Relacionados

Este capítulo revisa os trabalhos correlatos que fundamentam e contextualizam a presente pesquisa, com ênfase em sua relevância no contexto CPSs, estando organizado em três seções. A Seção 3.1 aborda os fundamentos dos ataques adversários em IDSs, discutindo suas características, estratégias de implementação, impacto nos sistemas de segurança e os principais problemas em aberto. Na Seção 3.2, a análise é ampliada para explorar como esses ataques afetam, de maneira específica, os IDSs voltados para CPSs, destacando os desafios adicionais impostos pela complexidade e criticidade desses sistemas. Por fim, a Seção 3.3 apresenta uma discussão aprofundada dos pontos levantados ao longo do capítulo, enfatizando as contribuições deste trabalho.

3.1 Ataques adversários em IDSs

Com o aumento da aplicação de AM em áreas de segurança, surgiram brechas que possibilitam a usuários mal-intencionados comprometer sistemas ao explorarem vulnerabilidades em algoritmos de AM. AYUB Md Ahsan; JOHNSON (2020) apresenta um estudo que explora o uso de ataques adversários para comprometer o funcionamento de IDSs baseados em algoritmos de AM. O algoritmo utilizado para o treinamento do IDS foi uma MLP. O cenário abordado considera que o atacante possui conhecimento sobre o tipo de modelo e seu funcionamento, caracterizando um ataque do tipo *White Box*. O ataque empregado foi o JSMA. Os autores constataram que, por meio desse ataque, o classificador sofre uma queda significativa de desempenho, demonstrando, portanto, vulnerabilidades contra ataques adversários em cenários onde há conhecimento prévio do modelo.

Yang et al. (2018) examina ataques adversários no contexto de NIDSs. O objetivo do estudo é investigar o desempenho de uma *Deep Neural Network* (DNN) treinada para identificar comportamentos anômalos em um modelo de *Black Box*. Nesse contexto, foram explorados três tipos de ataques: o ataque com modelo substituto utilizando o método CW para geração, o ataque baseado em *Zeroth-Order Optimization* (ZOO) e o ataque baseado

em *Generative Adversarial Network* (GAN). A base de dados utilizada foi o *Network Security Lab Knowledge Discovery Dataset* (NSL-KDD). O NSL-KDD, foi desenvolvido como uma versão aprimorada do *Knowledge Discovery and Data Mining 1999 Dataset* (KDD'99), o NSL-KDD foi apresentado por (TAVALLAEE et al., 2009). Os resultados demonstram que os ataques adversários são eficazes contra o classificador DNN neste cenário.

As DNNs têm demonstrado eficácia em diversas tarefas de AM, incluindo a detecção de intrusão. No entanto, estudos recentes apontam vulnerabilidades das DNNs a ataques adversários, especialmente no contexto da classificação de imagens. Esses ataques permitem que as redes sejam induzidas a gerar classificações incorretas com pequenas e imperceptíveis alterações nos pixels da imagem. Essa fragilidade levanta preocupações quanto ao uso de DNNs em áreas de segurança da informação, como no uso de IDSs. Entre os algoritmos de ataque adversário utilizados destacam-se o FGSM, JSMA, DeepFool e CW, sendo o classificador empregado uma MLP e a base de dados a NSL-KDD. Por fim, o estudo comprova a vulnerabilidade das DNNs frente a ataques adversários (WANG, 2018).

Pacheco e Sun (2021) destacam que pesquisas anteriores evidenciaram a vulnerabilidade dos algoritmos AM frente a ataques adversários em tarefas de classificação de imagens utilizando DNN. Contudo, o uso de IDSs carece de estudos mais abrangentes, uma vez que os esforços até o momento se concentraram principalmente nos conjuntos de dados KDD'99. Este conjunto de dados foi introduzido na competição KDD Cup 1999, pelos autores (STOLFO et al., 2000) e NSL-KDD. Este estudo propõe uma análise da suscetibilidade de conjuntos de dados contemporâneos, como o *UNSW Network-Based Dataset 15* (UNSW-NB15), esse conjunto de dados foi criado para abordar limitações dos conjuntos de dados anteriores, incorporando tráfego de rede moderno (MOUSTAFA; SLAY, 2015). Além disso o Bot-IoT, focado em ataques relacionados IoT, o BoT-IoT foi introduzido por (KORONIOTIS et al., 2019). Os ataques adversários examinados incluem o JSMA e o FGSM. Adicionalmente, a pesquisa confirma que as DNNs são efetivamente vulneráveis a ataques adversários.

Considerando as lacunas existentes, Li et al. (2020) fornecem um estudo abrangente sobre pesquisas que analisam ataques adversários e potenciais defesas em CPSs, utilizando diferentes tipos de dados de sensores, como dados de vigilância, áudio e texto. No decorrer da pesquisa, foi proposto um processo geral que inclui ataques adversários em CPSs, destacando possíveis estratégias de defesa para esses sistemas. Foi conduzida uma análise sistemática dos modelos de ataque e defesa existentes. Por fim, o estudo identificou lacunas e oportunidades para investigações futuras, como o aumento da eficácia dos exemplos adversários e a pesquisa de defesas que integrem múltiplos fatores (entradas, detectores, *back-end*, etc.) (LI et al., 2020).

Assim, Valeev (2022) oferecem uma análise estruturada sobre ataques adversários à

IA em CPSs, avaliando 45 artigos selecionados a partir de 134 publicações no banco de dados Scopus. O estudo abrange a categorização dos algoritmos de ataque e técnicas de defesa, além de apresentar uma perspectiva atualizada sobre as metodologias e ferramentas disponíveis. De forma similar, os autores identificam lacunas, como a distribuição desigual de publicações por domínio de aplicação de CPSs, a falta de padronização em metodologias de avaliação, o déficit de conjuntos de dados públicos de alta qualidade e o uso frequente de simulações no lugar de cenários do mundo real para avaliação.

Em Wang et al. (2023), é apresentado um estudo que resume os principais trabalhos sobre técnicas de ataque de evasão e defesa, delineando o estado da arte na segurança de modelos de AM em CPSs. Além disso, o estudo identifica algumas lacunas importantes. No que diz respeito aos ataques, destacam-se melhorias na transferibilidade e em ataques relacionados à alteração de conceitos adversários. Em contrapartida, as lacunas relacionadas às defesas incluem: conjuntos de dados preparados para ataques de evasão, ampliação do espaço de características, bloqueio da transferibilidade, fortalecimento do treinamento adversário e, por fim, a consideração simultânea das estratégias do atacante e do defensor, como a quantificação da percepção humana (WANG et al., 2023).

Alatwi e Aldweesh (2021) destacam que, além de serem vulneráveis a ataques adversários do tipo caixa branca, os métodos de AM também apresentam vulnerabilidades a ataques do tipo caixa preta, nos quais o atacante não possui informações prévias sobre o modelo. De forma abrangente, os autores oferecem um resumo sobre ataques adversários de caixa preta aplicados em IDSs, apontando seus principais desafios e questões em aberto.

Saranya et al. (2020) propõem um estudo sobre algoritmos de AM para a classificação do tráfego de rede como malicioso ou benigno. O método aplicado consistiu na implementação de diversos algoritmos de AM para avaliar sua eficácia na identificação de possíveis ataques. Além disso, os autores tiveram como objetivo comparar o desempenho dos algoritmos utilizados com base em métricas como acurácia, precisão, revocação e *F1-Score* (SARANYA et al., 2020).

De acordo com Quincozes et al. (2022), embora os métodos de detecção de intrusão sejam amplamente estudados em redes e sistemas convencionais, há uma escassez de pesquisas voltadas para os protocolos IEC-61850. Além disso, os autores revelam a existência de um gargalo na disponibilidade de dados industriais realistas para treinamento, teste e avaliação de IDSs. O estudo resultou em um *framework* para a geração de conjuntos de dados IEC-61850, com características específicas extraídas de protocolos de comunicação de subestação em nível de rede e do domínio elétrico, visando detectar diferentes tipos de ataques. A criação desse *framework* oferece uma solução para mitigar o gargalo de dados industriais, disponibilizando um conjunto de dados adequado para esse contexto.

3.2 Ataques adversários em IDSs para sistemas ciberfísicos

A crescente utilização de modelos de AM baseados em *Deep Learning* (DL) levanta preocupações sobre sua robustez e confiabilidade. Essa tecnologia é empregada em diversas áreas críticas, como *Industrial Cyber-Physical Systems*s (ICPSs), mas é suscetível a ataques adversários. Nesse contexto, a investigação a respeito desses ataques oferece informações valiosas para sua detecção e prevenção (GIPIŠKIS et al., 2023).

Khaw et al. (2024) destacam que a modernização dos sistemas de energia para SGs está proporcionando avanços significativos em confiabilidade, eficiência e percepção situacional. No entanto, essas melhorias também introduzem novas ameaças relacionadas à segurança da informação, devido ao grande volume de dados gerados pelas SGs e ao uso de AM como abordagem de segurança cibernética. Apesar dos benefícios, essa abordagem introduz brechas de segurança, especialmente no contexto de ataques adversários. O artigo propõe estudar a robustez de IDSs baseados em *autoencoders* em SGs contra ataques adversários. O ataque avaliado é um algoritmo de ataque adversário de evasão em um cenário de caixa branca. Os resultados sugerem que amostras adversárias iniciais não são eficazes para criar ataques adversários bem-sucedidos, evidenciando a complexidade de projetar ataques adversários contra IDSs baseados em *autoencoders* em SGs.

Anthi et al. (2021) estudam o uso de ataques adversários no contexto de SCIs e mencionam que seu trabalho é pioneiro ao investigar o comportamento de modelos supervisionados contra esses ataques. Além disso, os pesquisadores discutem como os sistemas SCIs podem se defender desses mecanismos de ataque. O trabalho faz uso da biblioteca em Python CleverHans[1] para gerar os ataques maliciosos. O ataque utilizado foi o JSMA, e os algoritmos testados foram RF e J48. Os resultados mostraram um decaimento no desempenho de ambos os algoritmos. No entanto, quando as amostras adversárias foram inseridas no treinamento, os classificadores demonstraram maior robustez em relação a tais ataques.

Figueroa, Wang e Giakos (2022) afirmam que algoritmos de AM e DL têm sido a vanguarda da IA, remodelando o cenário atual da computação. A proliferação da implantação desses modelos gera uma preocupação, pois eles são vulneráveis a ataques adversários. Dessa forma, ataques adversários podem afetar sistemas críticos, como os SCIs. Além disso, os SCIs utilizam algoritmos de AM para executar funções do dia a dia, fazendo com que esses sistemas se tornem alvos para ataques adversários. Assim, pesquisas sobre esses ataques podem melhorar a compreensão do assunto e evitar que usuários maliciosos explorem esses recursos de forma indevida. Para essa pesquisa, foram utilizados três ataques: FGSM, DeepFool e JSMA. Os resultados do estudo indicam que ataques adversários têm consequências negativas no desempenho de DL e algoritmos de AM em

[1] https://github.com/cleverhans-lab/cleverhans

SCIs.

O estudo de Silva, Miani e Zarpelao (2023) também abordou ataques adversários em cenários de SCIs. Foram analisados os ataques FGSM e JSMA, constatando-se que o conhecimento prévio sobre a estrutura do algoritmo-alvo pode intensificar a eficácia dos ataques. Outro fator relevante foi o volume de dados acessados pelo atacante, que também influenciou significativamente a intensidade dos ataques. Os resultados indicaram que, embora o FGSM tenha causado ataques com severidade média superior em comparação ao JSMA, este último se destaca por ser menos invasivo e, possivelmente, mais difícil de ser detectado (SILVA; MIANI; ZARPELAO, 2023). A Tabela 2 apresenta as características dos principais trabalhos relacionados.

3.3 Discussão

Com base nos trabalhos apresentados, é possível estabelecer uma discussão aprofundada relacionando suas contribuições e resultados com os objetivos do presente estudo.

- **Comparação entre contextos e metodologias:** A pesquisa de Gipiškis et al. (2023) explora os ICPSs e aponta a vulnerabilidade desses sistemas frente a ataques adversários, destacando a importância de medidas de prevenção e detecção. De forma similar, o presente estudo também aborda sistemas críticos, ampliando a análise para cenários com diferentes bases de dados e ataques. Enquanto o trabalho de Khaw et al. (2024) foca em SGs e IDSs baseados em *autoencoders*, nosso estudo utiliza múltiplos classificadores e metodologias, analisando o impacto em diferentes tipos de algoritmos e cenários de ataque, contribuindo com um panorama mais diversificado e aplicável.

- **Variedade de ataques adversários:** Trabalhos como os de Anthi et al. (2021) e Figueroa, Wang e Giakos (2022) analisam ataques específicos, como FGSM, JSMA e DeepFool, enquanto o estudo atual amplia essa abordagem para sete ataques distintos, incluindo os já mencionados. A análise de múltiplos cenários nos permite observar o comportamento dos IDSs em situações diversificadas, algo que ainda é uma lacuna em estudos anteriores.

- **Influência do treinamento com amostras adversárias:** Conforme evidenciado por Anthi et al. (2021), a inserção de amostras adversárias no treinamento aumenta a robustez dos classificadores. Este é um ponto de destaque no presente estudo, que investiga a eficácia dessa abordagem em diferentes bases de dados. Nossos resultados corroboram essas descobertas, especialmente para classificadores do grupo *ensemble learning*, que apresentaram maior resistência em cenários onde amostras adversárias foram utilizadas no treinamento.

- ❑ **Impacto dos ataques e variáveis analisadas:** De acordo com Silva, Miani e Zarpelao (2023), o FGSM causa maior severidade em comparação ao JSMA, mas o JSMA é mais difícil de ser detectado. Nosso estudo confirma essa tendência, mas avança ao avaliar como essas diferenças se manifestam em cenários com menor quantidade de dados ou maior variação de parâmetros de ataque. A análise do impacto em bases distintas, como *Power System Database* e *Ereno*, evidencia como os IDSs respondem de maneira diferente dependendo do contexto e dos dados utilizados.

- ❑ **Relevância do estudo em comparação à literatura:** Enquanto trabalhos como os de Figueroa, Wang e Giakos (2022) identificam as vulnerabilidades de DL em SCIs, o presente estudo amplia essa investigação para IDSs baseados em AM, considerando tanto classificadores individuais quanto comitê de classificadores. Essa abordagem integrada, aliada à análise de duas bases de dados distintas, proporciona um panorama mais amplo e detalhado sobre as vulnerabilidades e resistências dos modelos frente a ataques adversários.

Concluindo, o presente trabalho avança na literatura ao combinar diferentes bases de dados, metodologias e classificadores, considerando a distinção entre classificadores individuais e comitê de classificadores para analisar o impacto de ataques adversários em IDSs que utilizam AM em ambientes CPSs. A Tabela 2 resume esses resultados.

Tabela 2 – Síntese dos Trabalhos Relacionados.

Referência	**Classificadores**	**Conjunto de dados**	**Fontes de dados do modelo alvo**	**Classificadores Individuais X comitê de Classificadores**	**Ataques**
Anthi et al. [2021]	Random Forest, J48	*Power System Smart Grid Monitoring Power*	Medições em sistemas de energia elétrica	Não	JSMA
Figueroa et al. [2022]	Deep Neural Networks	*Industrial Control System (ICS) dataset*	Medições em sistemas de energia elétrica	Não	FGSM, DF, JSMA
Da Silva et al. [2023]	Random Forest, J48, MLP, AD	*Power System Smart Grid Monitoring Power*	Medições em sistemas de energia elétrica	Não	FGSM, JSMA
Gipiškis et al. [2023]	U-Net	*CTI FoodTech Peach Dataset* e *Semantic Drone Dataset*	Medições em sistemas de energia elétrica	Não	DAG
Khaw et al. [2024]	Autoencoders	*IEEE 14-Bus Test System*	Medições em sistemas de energia elétrica	Não	IBEA
Este estudo	*Classificadores individuais* (SVM, KNN, AD) e *comitê de classificadores* (RF, GBM, AdaBoost, XGB)	*Power System Database*, *Ereno IEC-61850 Intrusion Dataset*	Dados de redes elétricas e comunicações industriais	Sim	FGSM e JSMA

Capítulo 4

Materiais e Métodos

Este capítulo detalha o método adotado para o treinamento dos classificadores, bem como a execução e avaliação dos ataques adversários. O trabalho foi estruturado em duas etapas: na primeira, os classificadores são treinados, testados e avaliados sem interferências externas, estabelecendo uma linha de base; na segunda, são inseridos os ataques adversários para analisar o comportamento dos classificadores frente a essas situações

O capítulo está organizado em seis seções: a Seção 4.1 apresenta uma visão geral dos materiais e métodos; a Seção 4.2 detalha a linha de base utilizada para validar os classificadores; a Seção 4.3 discute a execução dos ataques adversários e a avaliação de seu impacto no desempenho dos modelos; a Seção 4.4 realiza uma análise semelhante à anterior, mas com foco na comparação entre classificadores individuais e comitê de classificadores e seus respectivos desempenhos; a Seção 4.5 explora um cenário de ataque alternativo aos discutidos anteriormente; e, por fim, a Seção 4.6 introduz um novo cenário de ataque.

4.1 Visão geral

Foram consideradas duas bases de dados: a primeira representa o cenário de SCIs, e a segunda, cenários baseados nos protocolos *IEC-61850*. A escolha das bases de dados para treinamento e desenvolvimento de testes é crucial para a construção da pesquisa, pois influencia diretamente os resultados obtidos. As fontes de informação utilizadas foram: *Power System Smart Grid Monitoring Power*, mencionada em (ALATWI; ALDWEESH, 2021), e *ERENO IEC-61850 Intrusion Dataset*, desenvolvida em (QUINCOZES et al., 2024). O *Power System Smart Grid Monitoring Power* foi selecionado por sua ampla utilização na literatura, sendo uma base consolidada em diversos trabalhos, enquanto o *ERENO IEC-61850 Intrusion Dataset* foi escolhido por ser um dataset mais recente, que incorpora a metodologia de enriquecimento de *features*, permitindo uma comparação mais robusta entre as bases de dados, representando cenários de SCIs construídos de formas distintas.

O *Power System Smart Grid Monitoring Power* é um conjunto de dados referente a um sistema de energia, desenvolvido pela Mississippi State University e Oak Ridge National Laboratory em 15 de abril de 2014. Essa base de dados é composta por três conjuntos derivados de um conjunto inicial, que consiste em 15 subconjuntos, cada um com 37 eventos de sistema de energia. Esses eventos são divididos em oito eventos naturais, um evento sem ocorrência e 28 eventos de ataque. Além disso, a base de dados possui 129 atributos, sendo 116 resultantes da medição de 29 atributos, em que cada um é medido por quatro *Unidades de Medição Fasorial* (UMF), e 12 atributos dedicados a registros do painel de controle, Snort, e uma última coluna para classificar o evento. O conjunto de dados original consiste em 55.663 amostras maliciosas e 22.714 benignas.

A segunda base de dados selecionada, *ERENO IEC-61850 Intrusion Dataset*, consiste em sete conjuntos de dados, cada um cobrindo cenários de ataque específicos. Cada conjunto de dados possui 69 atributos e diferentes quantidades de ataques e amostras normais. No total, o conjunto de dados possui 79.991 amostras maliciosas e 19.998 amostras benignas. Cada exemplo do conjunto de dados corresponde a uma mensagem sobre um evento gerenciado pelo protocolo *Generic Object Oriented Substation Event* (GOOSE), utilizado para troca de mensagens entre os dispositivos de uma subestação elétrica. Um determinado evento (malicioso ou não) corresponde a múltiplas mensagens.

As duas bases de dados selecionadas contribuem para a construção de uma pesquisa mais próxima da realidade e estão alinhadas com os objetivos deste trabalho, que busca analisar ataques adversários em CPSs, considerando dois grupos de algoritmos de AM: comitê de classificadores e classificadores individuais.

A Tabela 3 apresenta um resumo das informações utilizadas no método. Ademais, as Tabelas 4 e 5 listam todos os atributos pertencentes à base de dados *Power System Smart Grid Monitoring Power*. Mais informações sobre esses atributos, bem como o link para *download* da base, podem ser encontradas em [1]. Por sua vez, a Tabela 6 fornece a relação completa de atributos da base de dados *ERENO IEC-61850 Intrusion Dataset*. Informações adicionais sobre esses atributos, assim como o link para *download*, estão disponíveis em [2].

Tabela 3 – Visão geral dos materiais e métodos

Conjunto de Dados	**Número de Instâncias após Balanceamento**	**Pré-processamento**	**Treino/Teste**	**Ataques Adversários**	**Parâmetros dos Ataques**
A (Power System Smart Grid Monitoring Power)	28.882	MinMaxScaler(fit_transform), RandomUnderSampler(fit_resample)	70/30	FGSM, JSMA	eps = 0.01 / theta e gama = 0.1
B (ERENO IEC-61850 Intrusion Dataset)	27.938	MinMaxScaler(fit_transform), RandomUnderSampler(fit_resample)	70/30	FGSM, JSMA	eps = 0.01 / theta e gama = 0.1

Os dados passaram pela fase de pré-processamento, onde valores nulos ou inválidos foram removidos. A coluna de classificação foi convertida para formato binário, onde o

[1] <https://www.kaggle.com/datasets/bachirbarika/power-system>
[2] <https://www.kaggle.com/datasets/sequincozes/ereno-iec61850-ids>

Tabela 4 – Atributos do conjunto de dados Power System Smart Grid Monitoring Power (Parte 1)

#	**Atributo**	#	**Atributo**
1	R1-PA1:VH	66	R3-PM4:I
2	R1-PM1:V	67	R3-PA5:IH
3	R1-PA2:VH	68	R3-PM5:I
4	R1-PM2:V	69	R3-PA6:IH
5	R1-PA3:VH	70	R3-PM6:I
6	R1-PM3:V	71	R3-PA7:VH
7	R1-PA4:IH	72	R3-PM7:V
8	R1-PM4:I	73	R3-PA8:VH
9	R1-PA5:IH	74	R3-PM8:V
10	R1-PM5:I	75	R3-PA9:VH
11	R1-PA6:IH	76	R3-PM9:V
12	R1-PM6:I	77	R3-PA10:IH
13	R1-PA7:VH	78	R3-PM10:I
14	R1-PM7:V	79	R3-PA11:IH
15	R1-PA8:VH	80	R3-PM11:I
16	R1-PM8:V	81	R3-PA12:IH
17	R1-PA9:VH	82	R3-PM12:I
18	R1-PM9:V	83	R3:F
19	R1-PA10:IH	84	R3:DF
20	R1-PM10:I	85	R3-PA:Z
21	R1-PA11:IH	86	R3-PA:ZH
22	R1-PM11:I	87	R3:S
23	R1-PA12:IH	88	R4-PA1:VH
24	R1-PM12:I	89	R4-PM1:V
25	R1:F	90	R4-PA2:VH
26	R1:DF	91	R4-PM2:V
27	R1-PA:Z	92	R4-PA3:VH
28	R1-PA:ZH	93	R4-PM3:V
29	R1:S	94	R4-PA4:IH
30	R2-PA1:VH	95	R4-PM4:I
31	R2-PM1:V	96	R4-PA5:IH
32	R2-PA2:VH	97	R4-PM5:I

valor zero representa tráfego de rede normal, e o valor um indica tráfego de rede malicioso. Durante o pré-processamento, foi utilizada a técnica de normalização Min-Max, que calcula os valores mínimos e máximos dos atributos e utiliza esses valores para reescalar os dados para um intervalo específico (geralmente entre 0 e 1). Para aplicar essa normalização, utilizou-se a função fit_transform, que faz parte da biblioteca Python scikit-learn. Essa função realiza duas operações: primeiro, calcula os valores mínimos e máximos dos dados (com o método fit), e em seguida reescala os valores com base nesses limites (com o método transform).

Tabela 5 – Atributos do conjunto de dados Power System Smart Grid Monitoring Power (Parte 2)

#	**Atributo**	#	**Atributo**
33	R2-PM2:V	98	R4-PA6:IH
34	R2-PA3:VH	99	R4-PM6:I
35	R2-PM3:V	100	R4-PA7:VH
36	R2-PA4:IH	101	R4-PM7:V
37	R2-PM4:I	102	R4-PA8:VH
38	R2-PA5:IH	103	R4-PM8:V
39	R2-PM5:I	104	R4-PA9:VH
40	R2-PA6:IH	105	R4-PM9:V
41	R2-PM6:I	106	R4-PA10:IH
42	R2-PA7:VH	107	R4-PM10:I
43	R2-PM7:V	108	R4-PA11:IH
44	R2-PA8:VH	109	R4-PM11:I
45	R2-PM8:V	110	R4-PA12:IH
46	R2-PA9:VH	111	R4-PM12:I
47	R2-PM9:V	112	R4:F
48	R2-PA10:IH	113	R4:DF
49	R2-PM10:I	114	R4-PA:Z
50	R2-PA11:IH	115	R4-PA:ZH
51	R2-PM11:I	116	R4:S
52	R2-PA12:IH	117	control_panel_log1
53	R2-PM12:I	118	control_panel_log2
54	R2:F	119	control_panel_log3
55	R2:DF	120	control_panel_log4
56	R2-PA:Z	121	relay1_log
57	R2-PA:ZH	122	relay2_log
58	R2:S	123	relay3_log
59	R3-PA1:VH	124	relay4_log
60	R3-PM1:V	125	snort_log1
61	R3-PA2:VH	126	snort_log2
62	R3-PM2:V	127	snort_log3
63	R3-PA3:VH	128	snort_log4
64	R3-PM3:V	129	marker

Após a normalização, foi realizado um balanceamento das amostras de treino para garantir uma distribuição equilibrada entre as classes. Para isso, utilizou-se a técnica de subamostragem aleatória (Random Under Sampling, RUS), implementada pelo método RandomUnderSampler da biblioteca imbalanced-learn do Python. Esse método reduz a quantidade de amostras da classe majoritária, de forma aleatória, para igualá-la à quantidade de amostras da classe minoritária. A aplicação foi feita utilizando o método fit_resample, que ajusta (fit) os dados de entrada e retorna os conjuntos balanceados de variáveis independentes (x_treino) e dependentes (y_treino). Essa abordagem foi escolhida para evitar que o modelo seja enviesado pela predominância de uma classe, garan-

Tabela 6 – Atributos do conjunto de dados ERENO IEC-61850 Intrusion Dataset.

#	**Atributo**	#	**Atributo**
1	Time	18	vsbBTrapAreaSum
2	isbA	19	vsbCTrapAreaSum
3	isbB	20	t
4	isbC	21	GooseTimestamp
5	vsbA	22	SqNum
6	vsbB	23	StNum
7	vsbC	24	cbStatus
8	isbARmsValue	25	gooseTimeAllowedtoLive
9	isbBRmsValue	26	confRev
10	isbCRmsValue	27	stDiff
11	vsbARmsValue	28	sqDiff
12	vsbBRmsValue	29	cbStatusDiff
13	vsbCRmsValue	30	timestampDiff
14	isbATrapAreaSum	31	tDiff
15	isbBTrapAreaSum	32	timeFromLastChange
16	isbCTrapAreaSum	33	delay
17	vsbATrapAreaSum	34	marker

tindo assim uma melhor capacidade de generalização e um desempenho mais equilibrado nas classificações.

Após a seleção dos conjuntos de dados, procedeu-se à divisão em conjuntos de treino e teste. Esse procedimento foi aplicado de forma idêntica às duas bases de dados, alocando 70% dos dados para treino e 30% para teste. A estratégia *hold-out* foi empregada, sendo amplamente utilizada em estudos relacionados, conforme demonstrado por Anthi et al. (2021), Silva, Miani e Zarpelao (2023) e Figueroa, Wang e Giakos (2022). Durante o treinamento, os dados foram balanceados por meio do método *RandomUnderSampler*, disponibilizado pelo pacote *imbalanced-learn* em Python. Após o balanceamento, o conjunto de dados *Power System Smart Grid Monitoring Power* passou a contar com 14.441 instâncias, distribuídas entre amostras benignas e maliciosas, enquanto o *ERENO IEC-61850 Intrusion Dataset* passou a conter 13.969 amostras benignas e maliciosas.

Os algoritmos foram selecionados com base em dois critérios principais: (i) a formação de dois grupos representativos, um de classificadores individuais e outro de comitê de classificadores, e (ii) sua relevância na literatura. Foram escolhidos sete algoritmos, sendo três classificadores individuais (SVM, KNN e AD) e quatro comitê de classificadores (RF, GBM, AdaBoost e XGB). A seleção desses algoritmos baseou-se em sua recorrência em estudos sobre o desenvolvimento de IDSs baseados em AM. A escolha pela distinção entre classificadores individuais e comitê de classificadores justifica-se pela escassez de comparações diretas entre essas abordagens no contexto de IDSs para CPSs, tornando essa análise especialmente relevante. O mesmo conjunto de dados foi utilizado com a biblioteca

scikit-learn[3] do Python, e todos os algoritmos foram executados com seus valores padrão.

Em seguida, foi definida uma linha de base (*baseline*) como ponto de partida para a análise do desempenho dos classificadores sem a interferência de ataques adversários. Posteriormente, esse desempenho foi reavaliado considerando a presença desses ataques. A linha de base consistiu no treinamento dos algoritmos pré-selecionados, abrangendo tanto classificadores individuais quanto comitê de classificadores. Após o treinamento, os algoritmos foram testados, e seu desempenho foi medido com base no *F1-Score*.

As amostras adversárias foram geradas utilizando a biblioteca Python *Adversarial Robustness Toolbox* (ART), que disponibiliza diversos tipos de ataques adversários. Para este estudo, foram selecionados o JSMA e o FGSM. Os parâmetros *Theta* e *Gamma* do ataque JSMA foram configurados em 0,1, enquanto, no FGSM, o parâmetro *eps* foi definido como 0,01. Nesse cenário, o mesmo conjunto de dados utilizado para treinar os algoritmos foi "roubado"por um indivíduo malicioso e utilizado para treinar uma MLP, caracterizando um cenário *gray box*. Posteriormente, os ataques FGSM e JSMA utilizaram a MLP previamente treinada para gerar amostras de teste manipuladas.

A etapa final envolveu um novo treinamento dos classificadores com as amostras adversárias geradas. Nesse processo, as amostras adulteradas pelos ataques adversários foram incorporadas ao treinamento com o objetivo de aumentar a resistência dos modelos. Os classificadores treinados foram então testados novamente com os dados modificados, permitindo avaliar quais algoritmos e grupos apresentaram maior desempenho.

As subseções a seguir detalham as etapas aplicadas em cada cenário, abrangendo as duas bases de dados analisadas. O Capítulo 5 apresenta os resultados dos experimentos, destacando o impacto das técnicas adotadas em cada conjunto de dados.

4.1.1 Cenário de ataque

O presente estudo simula um cenário *gray box*, no qual o usuário malicioso possui conhecimento parcial do sistema. Esse atacante tem acesso limitado ao conjunto de dados de treinamento utilizado pelo modelo, além de informações restritas sobre sua arquitetura, como o tipo de modelo e os padrões de saída, sem, no entanto, dispor de detalhes completos sobre pesos ou hiperparâmetros. O atacante pode operar remotamente, explorando vulnerabilidades por meio de APIs públicas ou endpoints mal configurados na internet, ou localmente, na intranet, aproveitando permissões inadequadas em servidores. Além disso, ele pode ser um funcionário com acesso limitado, mas estratégico, capaz de utilizar essas informações para benefício próprio ou para causar danos à empresa.

No cenário *gray box*, o atacante detém um conhecimento intermediário, o que lhe permite realizar ações estratégicas sem controle total do sistema. Para roubar amostras, ele pode utilizar técnicas como *membership inference*, explorar vulnerabilidades em logs,

[3] <https://scikit-learn.org/stable/>

backups ou permissões expostas, ou ainda aproveitar um cargo interno que lhe garanta acesso privilegiado. Essas ações destacam a necessidade de implementar controles rigorosos de acesso aos dados, monitorar continuamente o pipeline de aprendizado e adotar medidas que detectem e mitiguem a inserção de entradas adversárias em qualquer etapa do processo.

A Figura 7 ilustra o cenário *gray box*, no qual o atacante possui acesso parcial às informações, como a base de dados de treinamento e dados relacionados ao treinamento do modelo. Nesse contexto, o atacante pode ser, por exemplo, um funcionário de uma empresa terceirizada com acesso limitado, porém suficiente para explorar vulnerabilidades ou obter informações sensíveis.

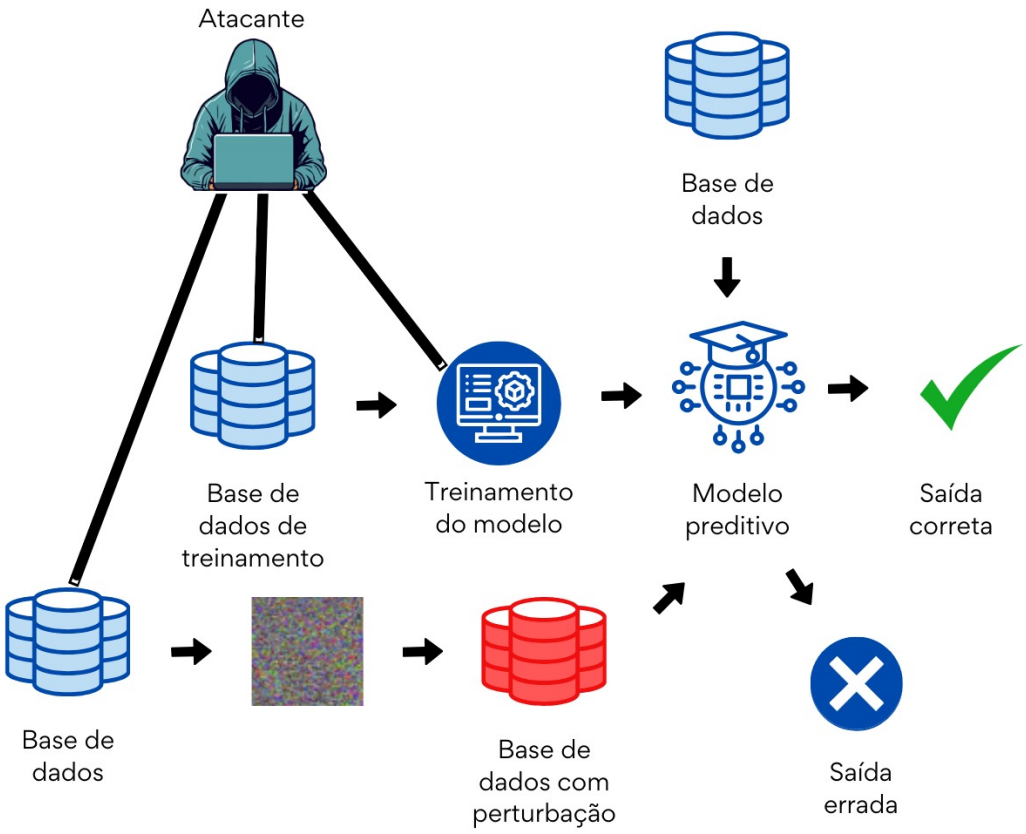

Figura 7 – Fluxo de um ataque adversário no cenário *gray box* aplicado a um IDS, ilustrando as interações limitadas entre o atacante e o modelo, com destaque para o tipo de evasão. Fonte: elaborado pelo autor (2024).

4.2 Estabelecimento da linha de base

A linha de base adotada nesta pesquisa compreende as seguintes sete etapas:

1. Seleção da base de dados associada a um CPS.
2. Realização do pré-processamento.
3. Separação da base de dados em 70% para treino e 30% para teste.
4. Seleção dos algoritmos:
   4.1 Escolha de um conjunto de algoritmos do tipo **classificadores individuais**.
   4.2 Escolha de um conjunto de algoritmos do tipo **comitê de classificadores**.
5. Treinamento do modelo de classificação binária para cada um dos algoritmos.
6. Teste do modelo de classificação binária para cada um dos algoritmos.
7. Avaliação do desempenho dos modelos usando a métrica *F1-Score*.

A Figura 8 ilustra as etapas mencionadas.

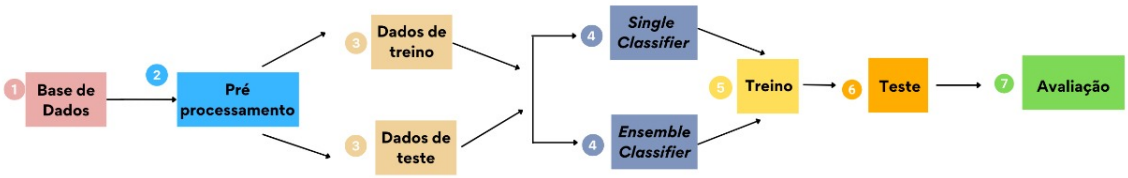

Figura 8 – Linha de base do método. Fonte: elaborado pelo autor (2024).

4.3 Comparação entre os ataques JSMA x FGSM

A hipótese H1 foi formulada para comparar os resultados obtidos na análise entre os ataques JSMA e FGSM. O objetivo foi verificar qual desses ataques exerce maior impacto no desempenho dos classificadores. Nesse contexto, a pesquisa foi conduzida em 11 etapas distintas, conforme descrito a seguir:

1. Seleção da base de dados associada a um CPS.
2. Realização do pré-processamento.
3. Separação da base de dados em 70% para treino e 30% para teste.
4. Seleção dos algoritmos:

4.1 Escolha de um conjunto de algoritmos do tipo classificadores individuais.

4.2 Escolha de um conjunto de algoritmos do tipo comitê de classificadores.

5. Treinamento do modelo de classificação binária para cada um dos algoritmos.
6. Teste do modelo de classificação binária para cada um dos algoritmos.
7. Avaliação do desempenho dos modelos.
8. Roubo do conjunto de treino (70%) e treinamento da MLP.
9. Modificação das amostras de ataque do conjunto de testes por meio dos ataques FGSM e JSMA, utilizando a MLP pré-treinada.
10. Apresentação do conjunto de testes modificado aos modelos de classificação treinados.
11. Reavaliação do desempenho dos modelos e comparação dos impactos dos ataques.

Todos os cenários de ataque podem ser considerados *Gray Box*, uma vez que o atacante tem apenas informações parciais a respeito dos dados, usando apenas o conjunto de treinamento. A Figura 9 ilustra as etapas.

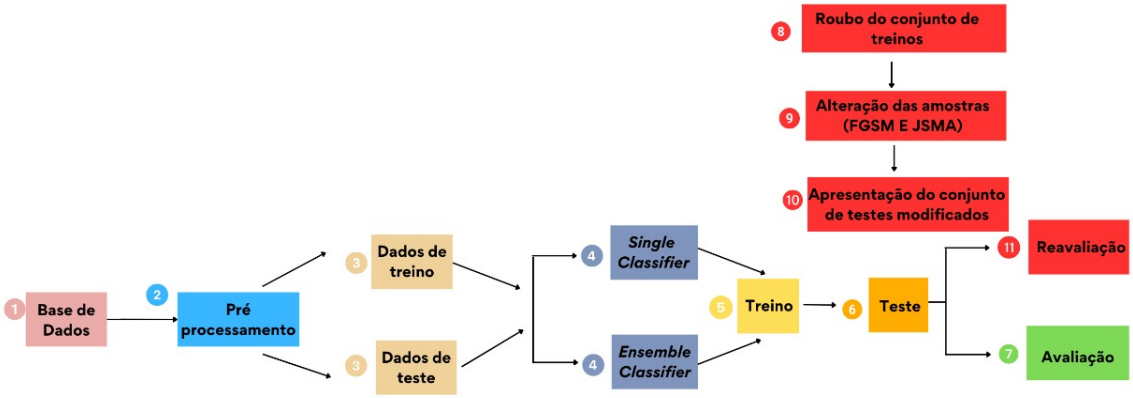

Figura 9 – Comparação do ataque JSMA x FGSM e avaliação de diferentes classes de algoritmos. Fonte: elaborado pelo autor (2024).

4.3.1 Variação dos parâmetros do JSMA e FGSM

Os passos seguidos neste contexto são os mesmos descritos na Seção 4.3, com exceção da etapa 9, que foi adaptada para incluir a variação dos parâmetros *theta* e *gama* do ataque JSMA, assim como o *eps* do ataque FGSM, visando avaliar o impacto dessas variações no desempenho dos modelos de detecção de intrusão.

9. Alteração das amostras de ataque do conjunto de testes utilizando dois ataques (FGSM e JSMA), em conjunto com a MLP previamente treinada. Os parâmetros *theta* e *gamma* do ataque JSMA, assim como o parâmetro *eps* do ataque FGSM, foram variados. Para o ataque JSMA, foi construída uma matriz, variando os valores de *theta* entre 0.1 e 0.9 e *gamma* entre 0.1 e 0.5, cruzando essas combinações para identificar aquela que causa o maior impacto negativo nos classificadores. No caso do ataque FGSM, o parâmetro *eps* foi ajustado entre 0.01 e 0.09, com o objetivo de determinar o valor que resulta no maior prejuízo ao desempenho dos classificadores.

As variações têm como objetivo identificar as combinações que causam os maiores impactos negativos nas métricas de desempenho dos classificadores. O parâmetro *theta*, no ataque adversário JSMA, define a magnitude da alteração aplicada a cada característica individual, representando a intensidade da perturbação inserida em atributos específicos dos dados. O parâmetro *gamma*, também associado ao JSMA, determina a proporção máxima de características que podem ser modificadas durante o ataque, ou seja, o limite percentual de atributos passíveis de perturbação em uma amostra. Por outro lado, o parâmetro *eps*, utilizado no ataque FGSM, especifica o valor máximo permitido para a perturbação adicionada a cada ponto dos dados de entrada, atuando como uma restrição à amplitude da modificação introduzida.

Assim, valores maiores de *eps* indicam ataques mais intensos em termos de alterações aplicadas, mas que também podem aumentar a probabilidade de detecção pelo IDS, uma vez que as mudanças nos dados se tornam mais perceptíveis.

Para esse teste, foram utilizados dois classificadores: o RF, representando o grupo comitê de classificadores, e a AD, representando o grupo classificadores individuais. Ambos foram selecionados por se destacarem dentro de seus respectivos grupos, conforme apresentado no Capítulo 5.

4.4 Comparação entre as classes de algoritmos (classificadores individuais e comitê de classificadores)

A hipótese H2 teve como objetivo avaliar os resultados da comparação entre as classes de algoritmos classificadores individuais e comitê de classificadores. O propósito foi determinar qual grupo apresenta melhor desempenho, ou seja, maior resiliência em IDSs baseados em anomalias em cenários de CPSs. As etapas utilizadas nesse cenário são semelhantes às descritas na Seção 4.3, com uma modificação específica apresentada a seguir:

11. Reavaliação do desempenho dos modelos e comparação entre os ataques. Nesta etapa, a reavaliação considera o desempenho dos classificadores, analisando qual dos grupos classificadores individuais ou comitê de classificadores demonstrou maior resiliência quando exposto aos ataques.

4.5 Diminuição da quantidade de amostras de treino que foram roubadas

A hipótese H3, por sua vez, buscou analisar se a redução na quantidade de amostras de treino roubadas contribui para diminuir significativamente a vulnerabilidade do modelo a ataques adversários. Esse ajuste visa tornar os modelos mais robustos para IDSs baseados em anomalias em cenários de CPSs.

Seguindo a mesma lógica da Seção 4.3, as etapas deste cenário são similares. No entanto, a etapa a seguir apresenta uma modificação específica:

8. Roubo do conjunto de treino (70%) e treinamento da MLP. Esse valor é reduzido pela metade, ou seja, o conjunto de treino representava 70% da base de dados e, desses 70%, foram capturados aleatoriamente 50%. Sendo assim a quantidade de amostras roubadas foi de 14.441 para o conjunto *Power System Smart Grid Monitoring Power* e de 13.969 para o conjunto *Ereno IEC-61850 Intrusion Dataset*.

Para esse processo, foi utilizada a função 'train_test_split' do Python, encontrada no módulo 'sklearn.model_selection'. Essa função divide dados em conjuntos de treinamento e teste, garantindo que cada conjunto seja uma amostra representativa para validação de modelos/classificadores. Ela aceita parâmetros como a proporção de divisão e randomização dos dados. A Figura 10 ilustra o processo.

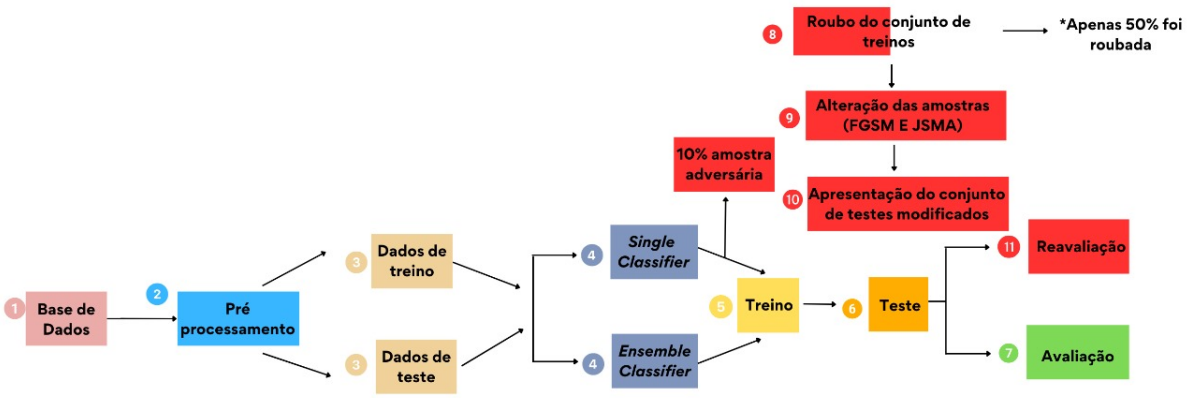

Figura 10 – Processo de diminuição da quantidade de amostras de treino que foram roubadas. Fonte: elaborado pelo autor (2024).

4.6 Inserção de amostras adversárias no conjunto de treino

A hipótese H4 teve como objetivo verificar se a inserção de amostras adversárias no conjunto de treino contribui para aumentar a robustez do modelo contra novos ataques, ampliando sua capacidade de generalização em cenários adversos. As etapas realizadas nesse cenário seguem, em sua maioria, as descritas na Seção 4.3, com a inclusão de uma nova etapa entre as etapas 4 e 5:

4. Seleção dos algoritmos.

4.5 Modificação do conjunto de treino: 10% aleatórios das amostras rotuladas como ataque foram substituídas por amostras adversárias aleatoriamente. Dois experimentos foram realizados: no primeiro, os 10% de amostras adversárias foram geradas pelo ataque JSMA, enquanto, no segundo, as amostras adversárias foram geradas pelo ataque FGSM. Dessa forma, foram gerados dois cenários distintos: o primeiro com 10% de amostras adversárias geradas pelo ataque JSMA, e o segundo com 10% de amostras adversárias geradas pelo ataque FGSM.

5. Treinamento do modelo de classificação binário para cada um dos algoritmos.

A Figura 11 demonstra esse cenário.

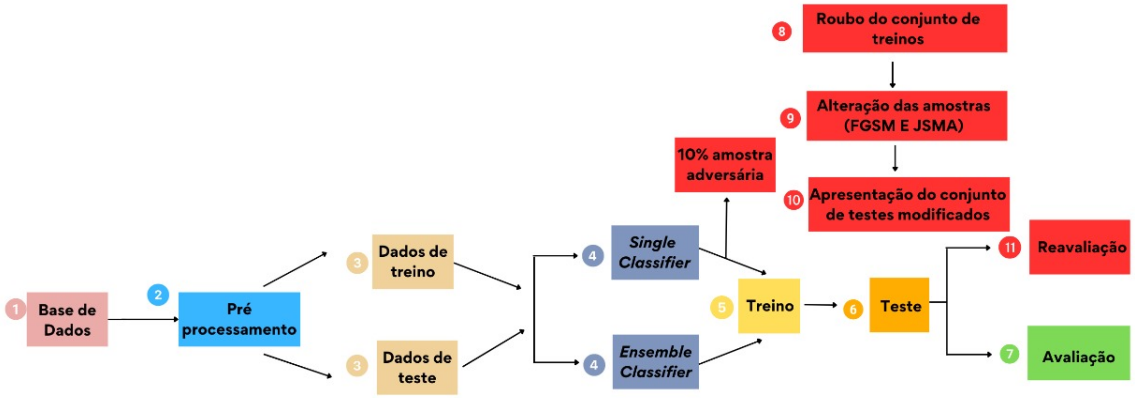

Figura 11 – Processo de inserção de amostras adversárias no conjunto de treino. Fonte: elaborado pelo autor (2024).

Capítulo 5

Análise dos Resultados

Neste capítulo, são apresentados os experimentos realizados, conforme descrito no Capítulo 4, bem como os resultados obtidos. O capítulo está organizado em seis seções, descritas a seguir: a Seção 5.1 apresenta os resultados detalhados da linha de base, utilizados como referência para a avaliação do desempenho dos classificadores; a Seção 5.2 discute os resultados dos ataques adversários e compara o desempenho entre eles; a Seção 5.3 examina os resultados obtidos pelos grupos classificadores individuais e comitê de classificadores, bem como seus respectivos algoritmos; a Seção 5.4 expõe os resultados dos classificadores após serem submetidos a um cenário de ataque alternativo; a Seção 5.5 apresenta os resultados obtidos em um novo cenário de ataque; e, por fim, a Seção 5.6 oferece uma análise geral dos resultados obtidos.

Inicialmente, sete classificadores de AM foram avaliados neste estudo. Três desses algoritmos pertencem à categoria classificadores individuais: SVM, KNN e AD. Os outros quatro compõem a categoria comitê de classificadores: RF, GBM, AdaBoost e XGB. Esses algoritmos foram treinados e testados utilizando duas bases de dados: *Power System Smart Grid Monitoring Power* e *ERENO IEC-61850 Intrusion Dataset*. Posteriormente, os classificadores foram submetidos a ataques adversários utilizando dois métodos: JSMA e FGSM.

A métrica utilizada para medir o desempenho dos classificadores foi o F1-score, que combina a precisão e a revocação em uma única medida. Valores de F1-score próximos de um indicam melhor desempenho do classificador, enquanto valores próximos de zero indicam pior desempenho. A fórmula usada para calcular o F1-Score é: Eq 3.

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3}$$

A precisão avalia a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo. Sua fórmula é dada pela Eq. 4:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{4}$$

A revocação mede a proporção de casos positivos corretamente identificados em relação ao total de casos positivos reais. Sua fórmula é apresentada pela Eq. 5:

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{5}$$

Onde:

- **VP**: Verdadeiros Positivos, ou seja, o número de previsões positivas corretas.
- **FP**: Falsos Positivos, ou seja, o número de previsões positivas incorretas.
- **FN**: Falsos Negativos, ou seja, o número de casos positivos reais que não foram identificados corretamente pelo modelo.

5.1 Resultados da *linha de base*

Nesta seção, são apresentados os resultados relacionados à linha de base, que serve como ponto de partida para a avaliação do desempenho dos classificadores em condições normais, ou seja, sem a aplicação de ataques adversários. O objetivo dessa análise inicial é estabelecer um referencial consistente, permitindo que o desempenho dos classificadores, sem a influência de ataques adversários, seja comparado às condições analisadas nas seções subsequentes, onde amostras adversárias serão introduzidas. Os classificadores avaliados incluem tanto os classificadores individuais quanto o comitê de classificadores, utilizando as bases de dados *Power System Smart Grid Monitoring Power* e *Ereno IEC-61850 Intrusion Dataset*. Essa abordagem será mantida ao longo das demais seções deste capítulo.

A Figura 12 apresenta o desempenho dos algoritmos na linha de base, avaliado por meio do F1-score. Esse critério foi aplicado tanto neste cenário quanto nos seguintes para a análise dos classificadores. A base de dados utilizada é a *Power System Smart Grid Monitoring Power*. A Figura 13 exibe os resultados obtidos com a mesma metodologia, utilizando, no entanto, a base de dados*Ereno IEC-61850 Intrusion Dataset*.

A partir dos resultados apresentados na Figura 12, observa-se que o melhor desempenho entre os classificadores do grupo classificadores individuais foi obtido pela AD, com um F1-score de 85%, seguida pelo KNN, com 81%, e pelo SVM, com 64%. No grupo comitê de classificadores, o destaque foi o RF, que alcançou um F1-score de cerca de 90%. Além disso, o XGB obteve 83%, enquanto o GBM e o AdaBoost apresentaram resultados em torno de 70%. De modo geral, os resultados da linha de base indicam uma vantagem ligeira dos classificadores do grupo comitê de classificadores em comparação aos do grupo classificadores individuais.

A partir dos resultados apresentados na Figura 13, observa-se que o melhor desempenho no grupo classificadores individuais foi obtido pela AD, com um F1-score de 91%.

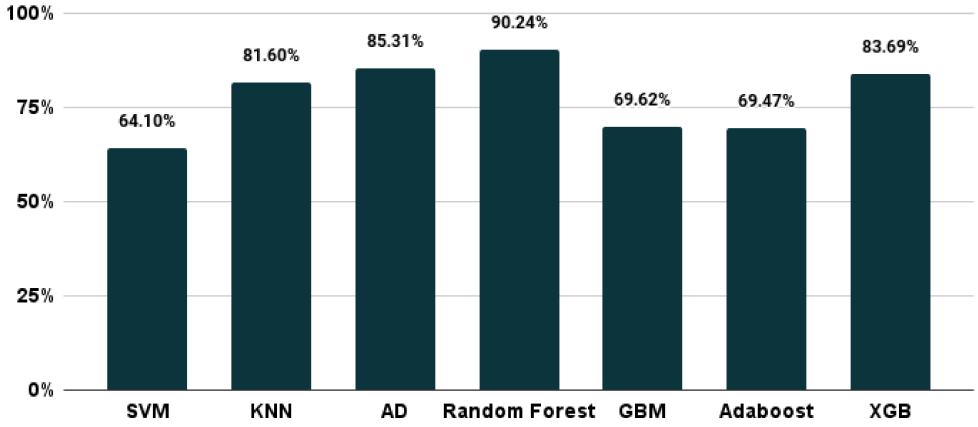

Figura 12 – Resultados da linha de base para o *Power System Smart Grid Monitoring Power*. Métrica: F1-Score. Fonte: elaborado pelo autor (2024).

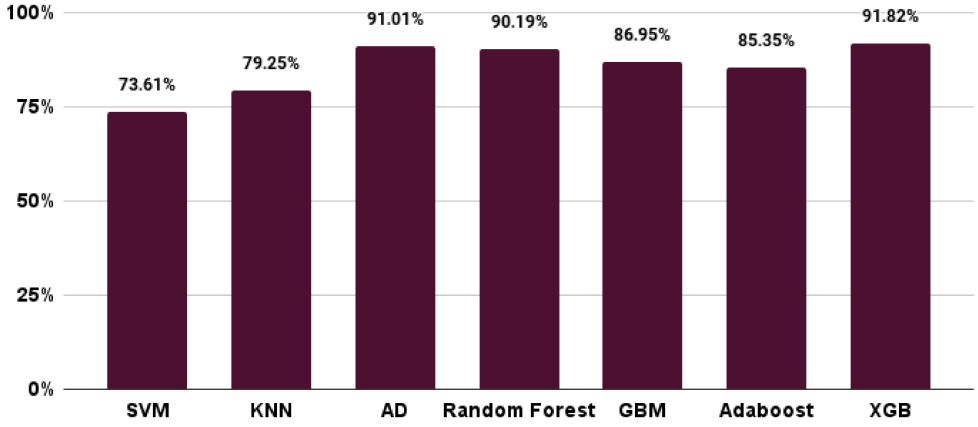

Figura 13 – Resultados da linha de base para o *Ereno IEC-61850 Intrusion Dataset*. Métrica: F1-Score. Fonte: elaborado pelo autor (2024).

No grupo comitê de classificadores, o destaque foi o XGB, com desempenho aproximado de 92%. Em comparação com a base anterior, os classificadores do grupo classificadores individuais apresentaram uma leve melhora; entretanto, os classificadores do grupo comitê de classificadores continuam demonstrando um desempenho marginalmente superior quando analisados de forma geral.

5.2 Comparação entre os ataques JSMA x FGSM

Nesta seção, é apresentada uma análise comparativa entre os ataques adversários FGSM e JSMA. O objetivo dessa comparação é avaliar a hipótese H1, ou seja, as diferenças no impacto de cada ataque sobre o desempenho dos classificadores. Para isso, é

analisado o F1-Score dos algoritmos.

Dois cenários são apresentados aqui: ataques FGSM e JSMA. A partir dos resultados obtidos na Figura 14, referente à base de dados *Power System Smart Grid Monitoring Power*, observa-se que, para ambos os grupos (classificadores individuais/comitê de classificadores), o ataque FGSM resulta em métricas inferiores ao JSMA. Esses resultados sugerem que o FGSM é aparentemente mais prejudicial aos classificadores testados, com exceção da AD. Além disso, nota-se que o grupo comitê de classificadores apresenta uma diferença mais acentuada no desempenho entre os dois ataques, enquanto o grupo classificadores individuais exibe variações menos significativas.

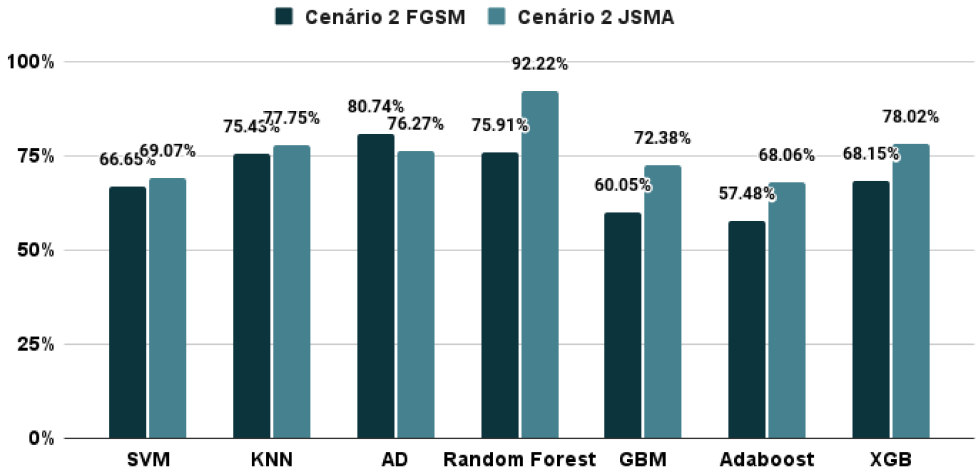

Figura 14 – Resultados da comparação entre os ataques JSMA x FGSM para o *Power System Smart Grid Monitoring Power*. Métrica: F1-Score. Fonte: elaborado pelo autor (2024).

Por outro lado, a Figura 15, referente à base de dados *ERENO IEC-61850 Intrusion Dataset*, revela que, nesse contexto, o grupo comitê de classificadores foi mais impactado pelo ataque JSMA. No grupo classificadores individuais, não foi identificado um padrão consistente: o classificador SVM apresentou maior impacto pelo ataque FGSM, o KNN mostrou maior vulnerabilidade ao ataque JSMA, enquanto a Árvore de Decisão exibiu resultados similares em ambos os ataques. Portanto, os resultados encontrados são consistentes com a hipótese H1. Como ambos os ataques possuem parâmetros que podem influenciar o desempenho do classificador, novos testes foram realizados, desta vez com a variação desses parâmetros. A Seção 5.2.1 detalha os resultados de tais testes.

5.2.1 Variação dos parâmetros do JSMA e FGSM

Agora o menor valor, 0.8997, na interseção de Theta = 0.7 Theta=0.7 e Gama = 0.5 Gama=0.5, está destacado em verde. Se precisar de mais ajustes, é só avisar!

As Tabelas 7 e 8, referentes à base de dados *Power System Smart Grid Monitoring Power*, ilustram os resultados gerados pela variação dos parâmetros *Theta/Gama* do

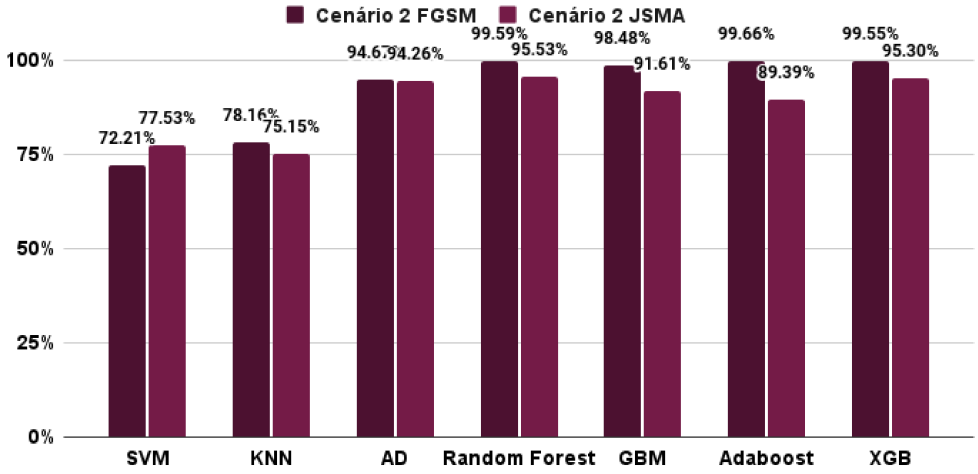

Figura 15 – Resultados da comparação entre os ataques JSMA x FGSM para o *Ereno IEC-61850 Intrusion Dataset*. Métrica: F1-Score. Fonte: elaborado pelo autor (2024).

ataque JSMA e seus respectivos impactos. A combinação de parâmetros que teve o maior impacto foi *Theta* = 0.7 e *Gama* = 0.5, o que resultou em um F1-score de 0,8997% para o classificador RF e 0,5792% para o classificador AD.

A Tabela 9 apresenta os resultados do ataque FGSM, considerando a variação do parâmetro *eps*. Os resultados indicam que o valor de *eps* mais efetivo para o RF foi 0.03, com um F1- Score de 0,7181%. Para o classificador AD o valor ideal de *eps* foi 0.06, resultando em um F1-Score de 0.5386%.

Tabela 7 – Resultados de F1-Score em porcentagem para *Random Forest* com parâmetros *Theta e Gama* variados. *Power System Smart Grid Monitoring Power*. Fonte: elaborado pelo autor (2024).

Random Forest						
	Gama	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Theta	0.1	0.9483	0.9358	0.9420	0.9402	0.9329
	0.2	0.9453	0.9415	0.9326	0.9297	0.9104
	0.3	0.9229	0.9441	0.9385	0.9445	0.9314
	0.4	0.9229	0.9398	0.9382	0.9447	0.9352
	0.5	0.9418	0.9016	0.9225	0.9409	0.9379
	0.6	0.9288	0.9359	0.9389	0.9410	0.9396
	0.7	0.9318	0.9232	0.9222	0.9452	0.8997
	0.8	0.9415	0.9395	0.9316	0.9465	0.9327
	0.9	0.9403	0.9456	0.9059	0.9459	0.9480

As Tabelas 10 e 11, referentes à base de dados *ERENO IEC-61850 Intrusion Dataset*, sumarizam os resultados gerados pela variação dos parâmetros *theta/gamma* do ataque JSMA e seus respectivos impactos. A combinação de parâmetros que teve o maior impacto no RF foi *theta* = 0.1 e *gamma* = 0.1, com um F1-Score de 0.9226%. Em relação à AD, a

Tabela 8 – Resultados de F1-Score em porcentagem para Árvore de Decisão com parâmetros *Theta e Gama* variados. *Power System Smart Grid Monitoring Power*. Fonte: elaborado pelo autor (2024).

Árvore de Decisão						
	Gama	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Theta	0.1	0.7833	0.7323	0.7882	0.7324	0.7564
	0.2	0.7374	0.7929	0.7104	0.7134	0.6507
	0.3	0.6727	0.7319	0.7419	0.7713	0.7650
	0.4	0.6766	0.7557	0.6557	0.7803	0.7562
	0.5	0.7539	0.6145	0.7083	0.7490	0.7435
	0.6	0.7527	0.7641	0.7753	0.7580	0.7386
	0.7	0.7146	0.6311	0.7278	0.7641	0.5792
	0.8	0.6569	0.7895	0.6944	0.7542	0.7400
	0.9	0.7495	0.7563	0.6435	0.6871	0.7802

Tabela 9 – Resultados de F1-Score em porcentagem para Random Forest e Árvore de Decisão com variação do parâmetro *eps* no ataque FGSM. *Power System Smart Grid Monitoring Power*. Fonte: elaborado pelo autor (2024).

eps	*Random Forest*	**Árvore de Decisão**
0.01	0.8706	0.7092
0.02	0.9226	0.7131
0.03	0.7181	0.6413
0.04	0.7675	0.7226
0.05	0.9316	0.8543
0.06	0.9245	0.5386
0.07	0.7998	0.6336
0.08	0.7239	0.6125
0.09	0.9582	0.7964

combinação mais impactante foi $theta = 0.1$ e $gamma = 0.5$, resultando em um F1-Score de 0.9275%.

A Tabela 12 apresenta os resultados do ataque FGSM, considerando a variação do parâmetro *eps*. Os resultados indicam que o valor de *eps* mais efetivo para o RF foi 0.03, resultando em um F1-Score de 0.9958%. Contudo, é importante notar que, neste cenário, o RF mostrou-se resistente aos ataques. Já para o classificador AD, o valor ideal de *eps* foi 0.09, mantendo um alto desempenho do classificador, com F1-Score de 0.9179%.

5.3 Comparação entre as classes de algoritmos classificadores individuais e comitê de classificadores

Nesta seção, é realizada uma comparação entre as classes de algoritmos de AM, com o objetivo de identificar as principais diferenças de desempenho em dois cenários distintos e validar a hipótese H2. O primeiro cenário representa as condições da linha de base, sem a

Tabela 10 – Resultados de F1-Score em porcentagem para Random Forest com parâmetros *Theta e Gama* variados. *ERENO IEC-61850 Intrusion Dataset*. Fonte: elaborado pelo autor (2024).

Random Forest						
	Gama	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Theta	0.1	0.9226	0.9385	0.9376	0.9807	0.9340
	0.2	0.9820	0.9951	0.9923	0.9873	0.9769
	0.3	0.9929	0.9901	0.9923	0.9780	0.9923
	0.4	0.9809	0.9887	0.9933	0.9953	0.9955
	0.5	0.9930	0.9569	0.9949	0.9842	0.9360
	0.6	0.9659	0.9962	0.9550	0.9943	0.9549
	0.7	0.9938	0.9894	0.9586	0.9931	0.9368
	0.8	0.9951	0.9901	0.9939	0.9956	0.9832
	0.9	0.9955	0.9889	0.9888	0.9802	0.9849

Tabela 11 – Resultados de F1-Score em porcentagem para Árvore de Decisão com parâmetros *Theta e Gama* variados. *ERENO IEC-61850 Intrusion Dataset*. Fonte: elaborado pelo autor (2024).

Árvore de Decisão						
	Gama	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Theta	0.1	0.9367	0.9404	0.9319	0.9726	0.9275
	0.2	0.9678	0.9807	0.9771	0.9735	0.9578
	0.3	0.9784	0.9758	0.9787	0.9599	0.9788
	0.4	0.9742	0.9702	0.9783	0.9807	0.9818
	0.5	0.9795	0.9420	0.9809	0.9721	0.9377
	0.6	0.9324	0.9813	0.9418	0.9781	0.9487
	0.7	0.9764	0.9730	0.9813	0.9785	0.9758
	0.8	0.9744	0.9776	0.9774	0.9816	0.9647
	0.9	0.9803	0.9744	0.9731	0.9647	0.9699

presença de ataques adversários. O segundo cenário é dividido em duas partes: a primeira avalia o impacto de ataques do tipo FGSM, e a segunda analisa os efeitos dos ataques do tipo JSMA. O objetivo dessa análise é evidenciar o desempenho dos classificadores nos cenários apresentados.

A Figura 16 apresenta os resultados dos algoritmos em dois cenários: o Cenário 1, no qual não há ataques adversários, e o Cenário 2, que considera os ataques adversários FGSM e JSMA. A base de dados utilizada é a *Power System Smart Grid Monitoring Power*. A Figura 17 aplica a mesma metodologia, porém utilizando a base de dados *ERENO IEC-61850 Intrusion Dataset*.

A partir dos resultados apresentados (Figura 16), observa-se que o melhor desempenho no Cenário 1 foi obtido pelo classificador RF, com aproximadamente 90%. A AD alcançou um desempenho próximo a 85%, seguida pelos classificadores XGB e KNN, que também se aproximaram de 85%, com o XGB ligeiramente superior ao KNN. Os demais algoritmos apresentaram resultados inferiores a 70%.

Tabela 12 – Resultados de F1-Score em porcentagem para o ataque FGSM variando o parâmetro *eps*. *ERENO IEC-61850 Intrusion Dataset*. Fonte: elaborado pelo autor (2024).

eps	*Random Forest*	**Árvore de Decisão**
=0.01	0.9961	0.9294
=0.02	0.9958	0.9504
=0.03	0.9958	0.9622
=0.04	0.9961	0.9381
=0.05	0.9960	0.9438
=0.06	0.9959	0.9356
=0.07	0.9961	0.9507
=0.08	0.9961	0.9531
=0.09	0.9958	0.9179

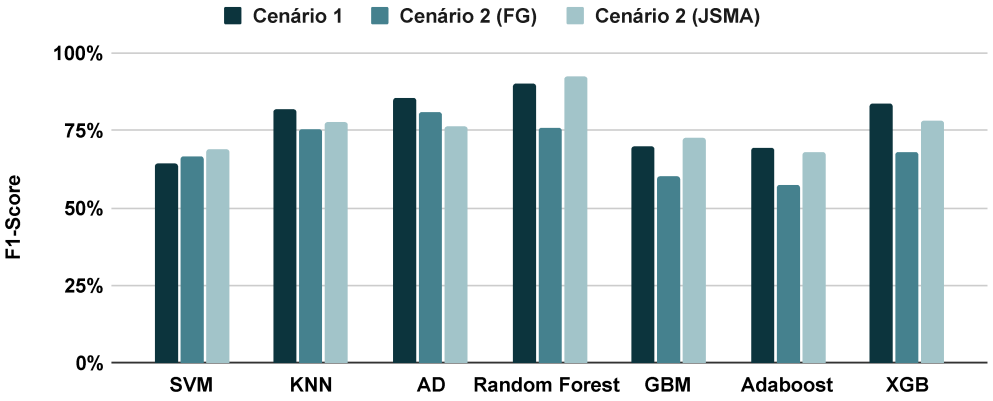

Figura 16 – Resultados para a comparação entre as classes de algoritmos(individual/comitê de classificadores). *Power System Smart Grid Monitoring Power*. Métrica: F1-Score Fonte: elaborado pelo autor (2024).

No Cenário 2, considerando o ataque FGSM, observa-se que o grupo classificadores individuais apresentou um menor declínio em seu desempenho em comparação ao grupo comitê de classificadores, com destaque para o classificador SVM, que, surpreendentemente, demonstrou uma melhora após o ataque. Além disso, a AD registrou o melhor F1-Score entre todos os classificadores após o ataque pelo FGSM.

No grupo comitê de classificadores, o padrão observado foi de um declínio considerável em todos os classificadores, embora o RF tenha mantido o melhor F1-Score dentro do grupo. Com exceção do SVM, os classificadores do grupo classificadores individuais também apresentaram uma queda em seus desempenhos.

No grupo classificadores individuais, o classificador que mais se destacou foi a AD, apresentando um F1-Score superior a 75% em todos os cenários e alcançando aproximadamente 80% no Cenário 1, onde não foi exposta a ataques adversários. Por outro lado, entre os classificadores do grupo comitê de classificadores, o algoritmo com melhor desempenho foi o RF, com F1-Scores superiores a 75% em todos os cenários e cerca de 90% no Cenário 1, sem ataques adversários. Curiosamente, o ataque JSMA resultou em

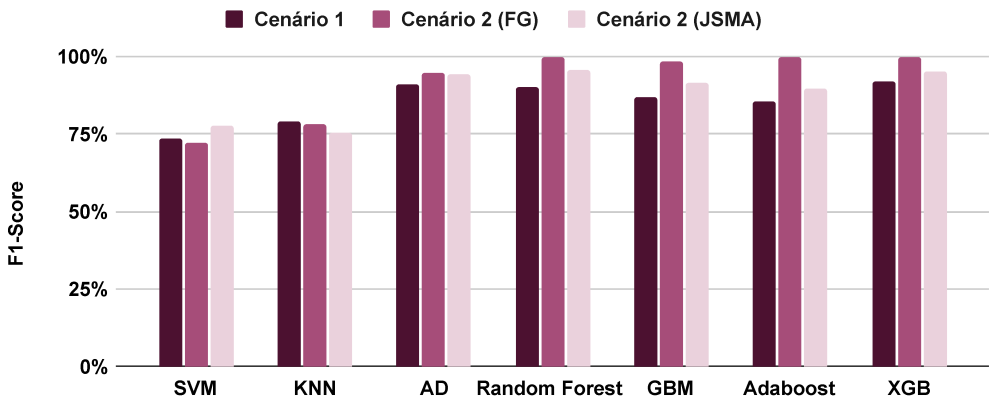

Figura 17 – Resultados para a comparação entre as classes de algoritmos(individual/comitê de classificadores). *Ereno IEC-61850 Intrusion Dataset*. Métrica: F1-Score Fonte: elaborado pelo autor (2024).

uma melhoria no F1-Score do RF, comportamento que merece destaque e análise mais detalhada.

Ainda no Cenário 2, mas considerando o ataque JSMA, observa-se que os algoritmos do grupo comitê de classificadores, RF e GBM, apresentaram um aumento em seus F1-Scores. Por outro lado, o AdaBoost teve uma leve queda no desempenho, enquanto o XGB registrou um decréscimo menor em comparação ao ataque FGSM. O RF obteve o melhor F1-Score entre todos os classificadores, considerando ambos os grupos, com um desempenho próximo de 95%. No grupo classificadores individuais, o SVM manteve o comportamento observado no ataque FGSM, ou seja, seu F1-Score aumentou; entretanto, com o JSMA, esse aumento foi ainda mais significativo. Além disso, o KNN apresentou um menor declínio quando exposto ao JSMA em comparação ao FGSM, enquanto a AD mostrou-se mais vulnerável ao JSMA do que ao FGSM.

A partir dos resultados obtidos na Figura 17, observa-se que os melhores classificadores de cada grupo, considerando o Cenário 1, foram o XGB e a AD, ambos com F1-Scores superiores a 90%. É importante destacar que o RF alcançou um desempenho quase equivalente ao do XGB. Os demais algoritmos do grupo comitê de classificadores registraram F1-Scores em torno de 85%, enquanto os classificadores individuais apresentaram F1-Scores próximos de 75%.

No Cenário 2, considerando o ataque FGSM, todos os classificadores do grupo comitê de classificadores apresentaram uma melhora em seus F1-Scores, atingindo aproximadamente 99%, demonstrando que não foram afetados pelo ataque. Por outro lado, no grupo classificadores individuais, o destaque foi novamente a AD, que também não foi impactada e alcançou um F1-Score de cerca de 95%. Os demais algoritmos foram afetados pelo ataque e tiveram desempenhos próximos de 75%.

Ainda no Cenário 2, mas considerando o ataque JSMA, os classificadores RF, XGB e AD se destacaram, alcançando F1-Scores próximos a 95%. O grupo comitê de classificado-

res, mais uma vez, não foi afetado pelo ataque e apresentou uma melhora nos F1-Scores de todos os seus classificadores. Em relação ao grupo classificadores individuais, tanto a AD quanto o SVM mantiveram seus desempenhos, sem apresentar depreciação no F1-Score. O KNN sofreu uma perda de desempenho maior em comparação ao ataque FGSM.

5.4 Diminuição da quantidade de amostras de treino que foram roubadas

Nesta seção, é analisado o impacto da diminuição da quantidade de amostras de treino roubadas pelo atacante, com o objetivo de avaliar a hipótese H3. Essa redução é aplicada como uma estratégia de mitigação para avaliar como a limitação de informações disponíveis no conjunto de treino influencia o desempenho dos ataques, considerando a métrica F1-Score dos classificadores. O objetivo principal é identificar quais algoritmos apresentam maior resiliência em cenários onde o atacante possui acesso restrito às amostras de treino, contribuindo para o desenvolvimento de CPS mais robustos e seguros.

A Figura 18 apresenta o desempenho dos algoritmos em três cenários distintos: no Cenário 1, não há ataques adversários; no Cenário 2, são introduzidos os ataques adversários FGSM e JSMA utilizando todos os dados de teste; e, por fim, no Cenário 3, também são aplicados ataques adversários FGSM e JSMA, porém o conjunto de treino roubado pelos atacantes é reduzido pela metade. A base de dados utilizada nesses experimentos é a *Power System Smart Grid Monitoring Power*. Já a Figura 19 ilustra os resultados da mesma metodologia, porém empregando a base de dados *Ereno IEC-61850 Intrusion Dataset*.

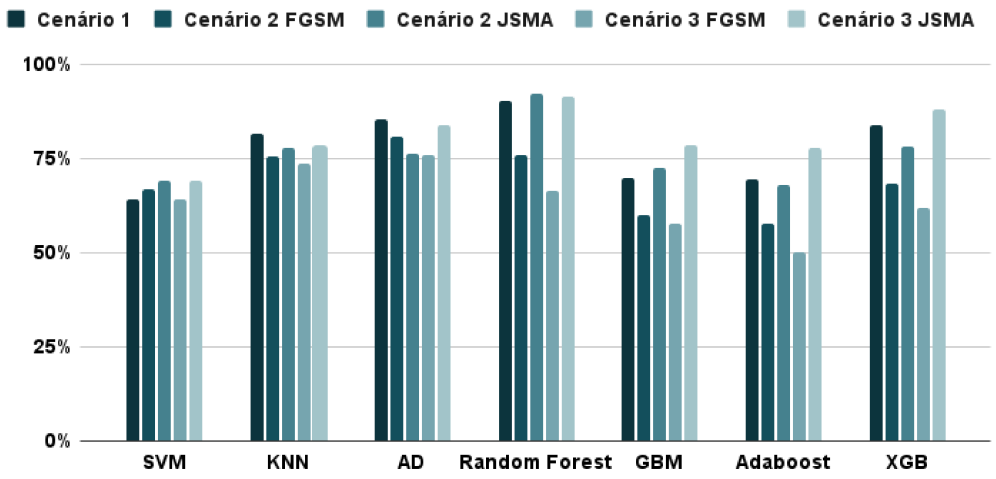

Figura 18 – Resultados para a diminuição da quantidade de amostras de treino que foram roubadas. *Power System Smart Grid Monitoring Power*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

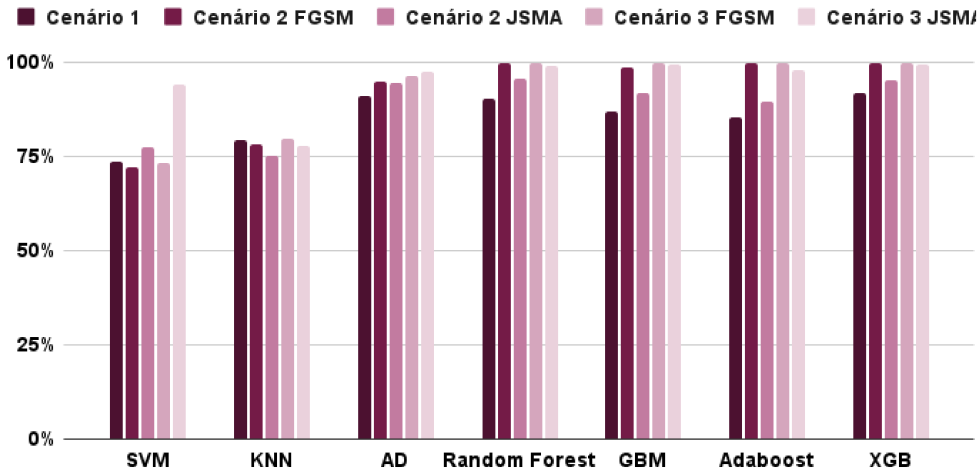

Figura 19 – Resultados para a diminuição da quantidade de amostras de treino que foram roubadas. *Ereno IEC-61850 Intrusion Dataset*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

A partir dos resultados obtidos (Figura 18), considerando o ataque FGSM no Cenário 3, observa-se que o melhor classificador foi a AD, com um desempenho de aproximadamente 75%. Em seguida, o KNN obteve um F1-Score próximo a 74%, enquanto os classificadores RF e SVM alcançaram desempenhos em torno de 67% e 65%, respectivamente. Os demais algoritmos apresentaram F1-Scores inferiores a 60%.

No Cenário 3, ainda considerando o ataque FGSM, observa-se que o grupo classificadores individuais apresentou uma redução de desempenho menos acentuada em comparação ao grupo comitê de classificadores, com destaque para a AD, que registrou o melhor F1-Score entre todos os classificadores após o ataque. Entre o grupo comitê de classificadores, verificou-se um impacto negativo considerável em todos os algoritmos, embora o RF tenha mantido o melhor F1-Score dentro desse grupo. Com exceção do SVM, os classificadores do grupo classificadores individuais também sofreram uma redução de desempenho; no entanto, essa queda foi menos significativa em relação ao grupo comitê de classificadores. Em resumo, a redução das amostras de treino roubadas pelo atacante durante o ataque FGSM tornou o ataque mais impactante para todos os classificadores, com exceção do SVM, sendo que o grupo comitê de classificadores apresentou uma queda de desempenho mais pronunciada.

No Cenário 3, considerando o ataque JSMA, observa-se que o grupo comitê de classificadores apresentou maior resistência em comparação ao grupo classificadores individuais. Todos os classificadores do grupo comitê de classificadores demonstraram desempenhos superiores após o ataque JSMA, quando comparados à linha de base. Os destaques desse grupo foram o RF e o XGB, ambos registrando F1-Scores de aproximadamente 90%. No grupo classificadores individuais, o melhor desempenho foi alcançado pela AD, que obteve um F1-Score de aproximadamente 84%. Em resumo, a redução das amostras de treino

roubadas pelo atacante durante o ataque JSMA reduziu a eficácia do ataque contra os classificadores testados na maioria dos cenários.

A partir dos resultados apresentados na Figura 19, considerando o ataque FGSM no Cenário 3, observa-se que os classificadores do grupo comitê de classificadores alcançaram os melhores desempenhos, com F1-Scores próximos a 99%. Em contraste, os classificadores do grupo classificadores individuais registraram F1-Scores em torno de 76%, com exceção da AD, que obteve um desempenho significativamente superior, com um F1-Score próximo a 96%.

No mesmo contexto, todos os classificadores do grupo comitê de classificadores apresentaram uma melhora em seus F1-Scores, alcançando aproximadamente 99%, e demonstraram não ser afetados pelo ataque. Por outro lado, no grupo classificadores individuais, o destaque foi novamente a AD, que não sofreu impacto e obteve um F1-Score de cerca de 95%. Entre os demais algoritmos, o KNN alcançou uma métrica superior a 75%, enquanto o SVM registrou um desempenho inferior a 75%.

No Cenário 3, considerando o ataque JSMA, o padrão observado foi semelhante ao do ataque FGSM. O grupo que se destacou foi o comitê de classificadores, com todos os classificadores alcançando F1-Scores próximos a 99%. Por outro lado, no grupo classificadores individuais, a AD obteve um F1-Score de aproximadamente 96%, enquanto o classificador SVM apresentou uma melhoria significativa em relação aos outros cenários, atingindo um F1-Score próximo a 95%. O classificador KNN, por sua vez, manteve um desempenho em torno de 75%.

Além disso, todos os classificadores do grupo comitê de classificadores apresentaram uma melhora em seus F1-Scores, alcançando aproximadamente 99%, e demonstraram não ser afetados pelo ataque. No grupo classificadores individuais, o destaque continuou sendo a AD, que manteve seu desempenho com um F1-Score de cerca de 95%. Nesta ocasião, o SVM também não foi impactado pelo ataque, enquanto o KNN sofreu uma redução em seu desempenho.

5.5 Inserção de amostras adversárias no conjunto de treino

Nesta seção, é analisado o impacto da inserção de amostras adversárias no conjunto de treino dos classificadores, com a intenção de validar a hipótese H4. Essa estratégia busca avaliar se a exposição dos algoritmos a dados adversários durante a fase de treinamento pode melhorar sua capacidade de generalização e resiliência frente a ataques futuros. O objetivo é identificar cenários em que a inclusão de amostras adversárias contribui para o aprimoramento do desempenho dos classificadores, fortalecendo sua robustez em sistemas ciberfísicos. Esta análise contempla os Cenários 1, 2 e 4.

As Figuras 20 e 21 apresentam os resultados dos algoritmos em três cenários distintos: no Cenário 1 (linha de base), não há ataques adversários; no Cenário 2, são acrescentados os ataques adversários FGSM e JSMA; e, finalmente, no Cenário 4, são inseridas amostras adversárias no conjunto de treino. O Cenário 4 foi subdividido em quatro etapas:

1. **Cenário 4 JSMA-FG:** Amostras do ataque JSMA são introduzidas no conjunto de treino, enquanto os classificadores são atacados pelo FGSM.
2. **Cenário 4 JSMA-JSMA:** Amostras do ataque JSMA são introduzidas no conjunto de treino, e os classificadores são atacados pelo próprio JSMA.
3. **Cenário 4 FG-FG:** Amostras do ataque FGSM são introduzidas no conjunto de treino, enquanto os classificadores são atacados pelo FGSM.
4. **Cenário 4 FG-JSMA:** Amostras do ataque FGSM são introduzidas no conjunto de treino, e os classificadores são atacados pelo JSMA.

A base de dados utilizada neste caso é a *Power System Smart Grid Monitoring Power*. A Figura 21 aplica a mesma metodologia, mas utilizando a base de dados *Ereno IEC-61850 Intrusion Dataset*.

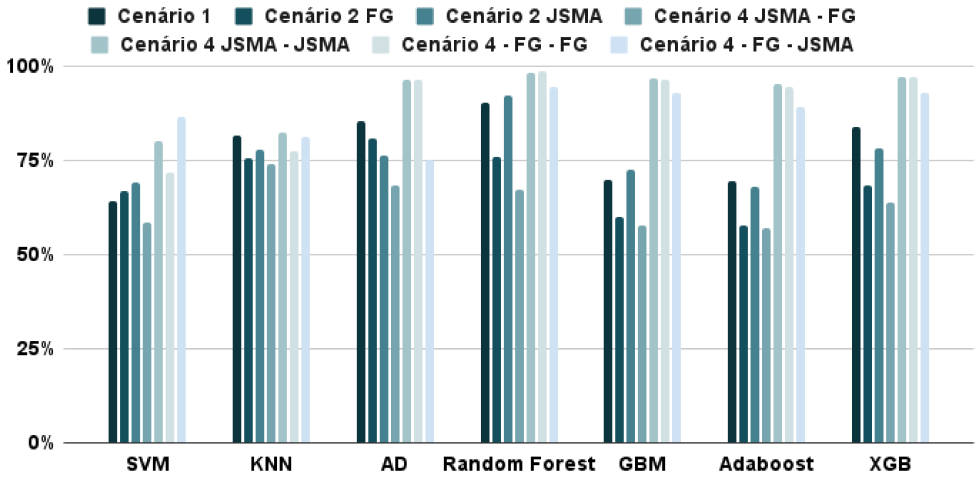

Figura 20 – Resultados para a inserção de amostras adversárias no conjunto de treino *Power System Smart Grid Monitoring Power*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

A partir das Figuras 20 e 21, observa-se que a inserção de amostras adversárias na fase de treinamento desempenhou um papel significativo como mecanismo de segurança contra ataques adversários para a maioria dos algoritmos avaliados. Em ambos os conjuntos de dados analisados (*Power System Smart Grid Monitoring Power* e *Ereno IEC-61850 Intrusion Dataset*), é possível observar que os classificadores demonstraram um desempenho superior em cenários onde as amostras adversárias foram incorporadas no treino.

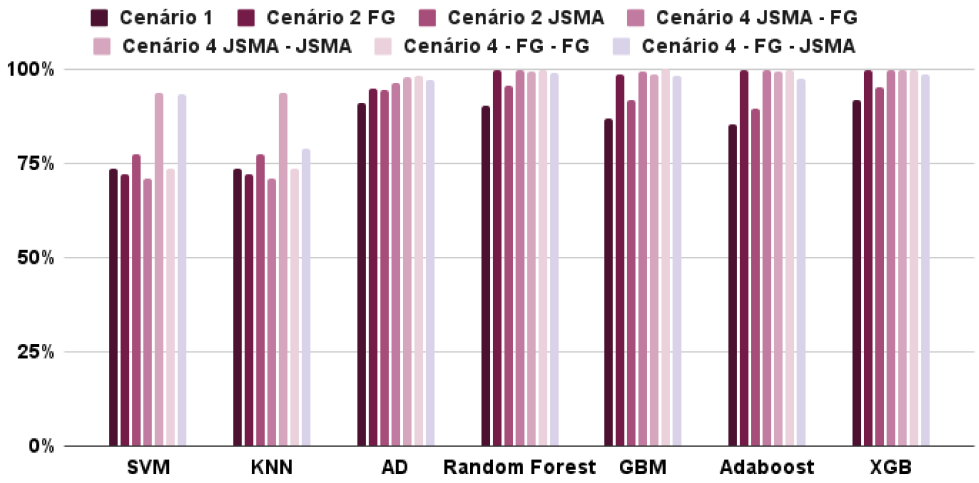

Figura 21 – Resultados para a inserção de amostras adversárias no conjunto de treino *Ereno IEC-61850 Intrusion Dataset*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

De acordo com a Figura 20, que leva em consideração a base de dados *Power System Smart Grid Monitoring Power*, é possível verificar que os resultados são melhores quando o conjunto de treino é exposto ao ataque FGSM, isso em ambos os casos, Cenário 4 FG-FG e Cenário 4 FG-JSMA. Porém, quando o conjunto de treino é exposto ao ataque JSMA o desempenho dos classificadores diminui no Cenário 4 JSMA-FG e no Cenário 4 JSMA-JSMA o desempenho volta a ser melhor do que a linha de base. Considerando a Figura 21 que leva em consideração a base de dados *Ereno IEC-61850 Intrusion Dataset*, é possível verificar que para o grupo classificadores individuais o Cenário 4 FG-FG aumentou o desempenho apenas do classificador Árvore de Decisão. Considerando o Cenário 4 FG-JSMA, é notado que o desempenho de todos os classificadores aumentou. Em relação ao Cenário 4 JSMA-FG apenas os classificadores SVM e KNN não melhoraram o seu desempenho. Por fim, no Cenário 4 JSMA-JSMA todos os classificadores melhoraram o desempenho.

Conforme ilustrado na Figura 20, verifica-se que a inserção de amostras adversárias geradas pelo ataque JSMA durante a fase de treinamento resultou em alterações substanciais no desempenho dos classificadores. Quando submetidos ao ataque FGSM, os grupos classificadores individuais e comitê de classificadores apresentaram os piores resultados entre todos os cenários analisados. Em contrapartida, a exposição ao próprio ataque JSMA proporcionou os melhores resultados em relação à linha de base, com exceção do classificador KNN, cujo desempenho permaneceu praticamente inalterado.

No Cenário 4 JSMA-FG, o KNN destacou-se como o classificador com o melhor desempenho, atingindo um F1-Score de aproximadamente 75%. Contudo, todos os algoritmos, tanto do grupo classificadores individuais quanto do grupo comitê de classificadores, apresentaram uma redução no desempenho em relação aos resultados obtidos na linha de base.

No Cenário 4 JSMA-JSMA, observa-se que o grupo comitê de classificadores registrou uma melhoria expressiva no desempenho, com todos os classificadores atingindo F1-Scores próximos a 98%. No grupo classificadores individuais, o destaque foi a AD, que alcançou um F1-Score em torno de 95%. Adicionalmente, o classificador SVM evidenciou um progresso considerável em relação à linha de base, com um aumento no F1-Score de aproximadamente 65% para cerca de 80%.

No Cenário 4 FG-FG, os algoritmos do grupo comitê de classificadores continuaram a demonstrar elevado desempenho, alcançando F1-Scores próximos a 95%. No grupo classificadores individuais, a AD manteve-se como o classificador com o melhor desempenho. O SVM apresentou, mais uma vez, uma melhoria substancial em relação à linha de base, enquanto o KNN registrou uma leve redução no desempenho em comparação à linha de base.

No contexto do Cenário 4 FG-JSMA, os algoritmos do grupo comitê de classificadores mantiveram sua relevância, embora os resultados tenham sido marginalmente inferiores em comparação ao cenário anterior. Por outro lado, a AD apresentou, neste caso, um desempenho inferior em relação à linha de base, alcançando um F1-Score de aproximadamente 75%. Dentro do grupo classificadores individuais, o SVM destacou-se, alcançando seu melhor desempenho em todos os cenários, com um F1-Score de cerca de 86%. O KNN também registrou uma melhoria significativa, atingindo um F1-Score de aproximadamente 81%.

Com base na Figura 21, é possível observar que o Cenário 4 JSMA-FG proporcionou ganhos de desempenho para os algoritmos do grupo comitê de classificadores em comparação à linha de base, com todos os classificadores deste grupo obtendo F1-Scores próximos a 99%. No grupo classificadores individuais, os algoritmos SVM e KNN evidenciaram uma leve redução em seus desempenhos, ambos com F1-Scores em torno de 71%. Por outro lado, a AD destacou-se no grupo classificadores individuais, alcançando um F1-Score de aproximadamente 96%.

No Cenário 4 JSMA-JSMA, todos os algoritmos apresentaram desempenhos superiores em relação à linha de base. O grupo comitê de classificadores obteve os melhores resultados, com F1-Scores próximos a 99%. No grupo classificadores individuais, os algoritmos SVM e KNN destacaram-se pela maior evolução em relação à linha de base, com ambos elevando seus F1-Scores de aproximadamente 75% para cerca de 93%.

No Cenário 4 FG-FG, os resultados observados para o grupo classificadores individuais foram consistentes com os do cenário anterior. A AD manteve elevado desempenho, com um F1-Score aproximando-se de 98%. Em contrapartida, os algoritmos SVM e KNN apresentaram desempenhos semelhantes aos da linha de base, ambos registrando F1-Scores em torno de 73%.

Ainda no Cenário 4 FG-FG, os resultados para o grupo classificadores individuais foram consistentes com os observados no cenário anterior. A AD manteve seu desempenho

elevado, com F1-Score aproximando-se de 98%. No entanto, o SVM apresentou uma melhoria considerável, elevando seu F1-Score de aproximadamente 73% na linha de base para cerca de 93%. Por sua vez, o KNN registrou um incremento mais modesto, com o F1-Score subindo de aproximadamente 73% para cerca de 79%.

5.6 Discussão

Nesta seção, são discutidos os resultados obtidos nos experimentos realizados. A análise busca relacionar os resultados com os objetivos propostos, avaliando as implicações práticas das descobertas para o desenvolvimento de IDSs mais robustos e eficazes em CPSs. Além disso, são destacadas as limitações identificadas ao longo do estudo. Essa discussão permite contextualizar os resultados, enfatizando a importância do presente estudo diante dos desafios existentes na literatura.

As Figuras 22 e 23 representam, respectivamente, a base de dados *Power System Smart Grid Monitoring Power* e a base de dados *ERENO IEC-61850 Intrusion Dataset*, incluindo todos os cenários agrupados. Os resultados indicam que os classificadores apresentaram melhor desempenho quando treinados e testados com a base de dados *ERENO IEC-61850 Intrusion Dataset*. A Tabela 13 sumariza todos os cenários analisados e os respectivos objetivos de cada um.

Tabela 13 – Principais resultados de todos os cenários apresentados nos resultados. Fonte: elaborado pelo autor (2024).

Cenário	Dataset	Ataque	F1-Score Médio (%)	Melhor Desempenho	Pior Desempenho
Cenário 1	Power System	Sem Ataque	Single: 77.00 ± 9.29, Ensemble: 78.25 ± 9.01	AD, RF	SVM, Adaboost
	Ereno IEC	Sem Ataque	Single: 79.41 ± 7.54, Ensemble: 88.57 ± 2.55	AD, XGB	SVM, Adaboost
Cenário 2	Power System	FGSM	Single: 74.27 ± 5.80, Ensemble: 65.39 ± 7.23	AD, RF	SVM, AdaBoost
	Ereno IEC	FGSM	Single: 79.03 ± 10.49, Ensemble: 99.32 ± 0.48	AD, Adaboost	SVM, GBM
Cenário 2	Power System	JSMA	Single: 74.36 ± 3.78, Ensemble: 77.67 ± 9.11	AD, KNN	SVM, GBM
	Ereno IEC	JSMA	Single: 83.10 ± 7.88, Ensemble: 92.96 ± 2.68	AD, RF	SVM, Adaboost
Cenário 3	Power System	FGSM	Single: 71.18 ± 5.07, Ensemble: 58,95 ± 5.87	AD, RF	SVM, AdaBoost
	Ereno IEC	FGSM	Single: 80.27 ± 11.23, Ensemble: 99.65 ± 0.04	AD, AdaBoost	SVM, XGB
Cenário 3	Power System	JSMA	Single: 76.90 ± 6.02, Ensemble: 80.00 ± 2.10	AD, RF	SVM, Adaboost
	Ereno IEC	JSMA	Single: 95.94 ± 3.23, Ensemble: 98.85 ± 0.54	AD, GBM	KNN, Adaboost
Cenário 4 FG-FG	Power System	FGSM	Single: 81.81 ± 10.54, Ensemble: 96.64 ± 1.42	AD, RF	SVM, Adaboost
	Ereno IEC	FGSM	Single: 81.73 ± 10.89, Ensemble: 99.74 ± 0.11	AD, GBM	KNN, RF
Cenário 4 FG-JSMA	Power System	JSMA	Single: 89.83 ± 7.39, Ensemble: 98.28 ± 0.53	SVM, RF	AD, Adaboost
	Ereno IEC	JSMA	Single: 86.00 ± 2.00, Ensemble: 91.00 ± 1.20	AD, RF	KNN, Adaboost
Cenário 4 JSMA-FG	Power System	FGSM	Single: 66.95 ± 6.36, Ensemble: 61.35 ± 4.23	AD, RF	SVM, AdaBoost
	Ereno IEC	FGSM	Single: 79.53 ± 10.52, Ensemble: 99.60 ± 0.18	AD, Adaboost	SVM, GBM
Cenário 4 JSMA-JSMA	Power System	JSMA	Single: 86.11 ± 7.20, Ensemble: 96.74 ± 1.18	AD, RF	SVM, Adaboost
	Ereno IEC	JSMA	Single: 94.29 ± 2.00, Ensemble: 99.16 ± 0.56	AD, XGB	KNN, Adaboost

A partir da Tabela 13, é possível observar os principais resultados:

- **Power System Smart Grid Monitoring Power (cenários com os piores resultados):**
  - **Cenário 3 (FGSM):** O desempenho médio mais baixo foi observado neste cenário, com o grupo classificadores individuais apresentando uma média de 71.18% ± 5.07 e o grupo comitê de classificadores com 58.95% ± 5.87.

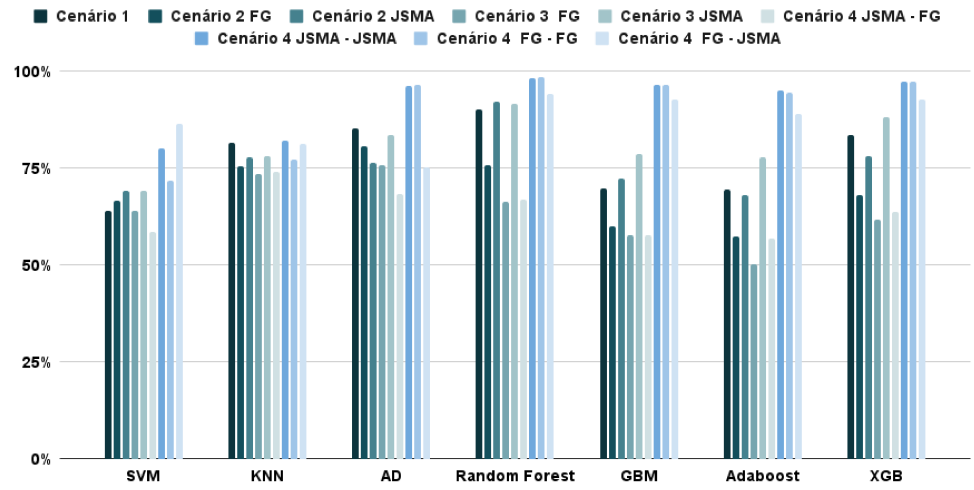

Figura 22 – Compilado dos resultados para o *Power System Smart Grid Monitoring Power*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

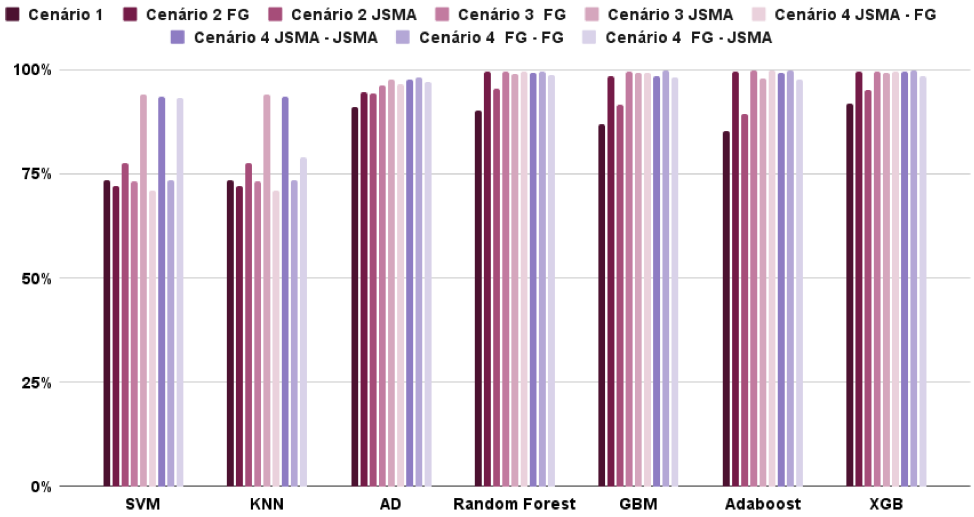

Figura 23 – Compilado dos resultados para o *Ereno IEC-61850 Intrusion Dataset*. Métrica F1-Score. Fonte: elaborado pelo autor (2024).

- **Cenário 4 JSMA-FG:** Este cenário também apresentou baixa performance, com o grupo classificadores individuais alcançando 66.95% ± 6.36 e o grupo comitê de classificadores ficando em 61.35% ± 4.23.

❑ **Ereno IEC-61850 Intrusion Dataset (cenários com os piores resultados):**

- **Cenário 4 JSMA-FG:** O desempenho mais baixo foi observado aqui, com o grupo classificadores individuais alcançando 79.53% ± 10.52 e o grupo comitê de classificadores atingindo 99.60% ± 0.18.
- **Cenário 3 (FGSM):** Outro cenário com baixo desempenho, onde o grupo classificadores individuais obteve 80.27% ± 11.23 e o grupo comitê de classifi-

cadores 99.65% ± 0.04.

- ***Power System Smart Grid Monitoring Power* (cenários com os melhores resultados):**
  - **Cenário 4 JSMA-JSMA:** Este cenário apresentou o melhor desempenho, com o grupo classificadores individuais alcançando um F1-Score médio de 86.11% ± 7.20 e o grupo comitê de classificadores atingindo 96.74% ± 1.18. Os melhores classificadores foram o AD no grupo classificadores individuais e o RF no grupo comitê de classificadores.
  - **Cenário 4 FG-FG:** Também apresentou alta resistência, com o grupo classificadores individuais registrando um F1-Score médio de 81.81% ± 10.54 e o grupo comitê de classificadores alcançando 96.64% ± 1.42. Os melhores desempenhos foram do AD no grupo classificadores individuais e do RF no grupo comitê de classificadores.
- ***Ereno IEC-61850 Intrusion Dataset (cenários com os melhores resultados):***
  - **Cenário 4 FG-FG:** Este cenário apresentou a maior resistência, com o grupo classificadores individuais atingindo um F1-Score médio de 81.73% ± 10.89 e o grupo comitê de classificadores alcançando 99.74% ± 0.11. Os melhores desempenhos foram do AD no grupo classificadores individuais e do GBM no grupo comitê de classificadores.
  - **Cenário 4 JSMA-JSMA:** Apresentou desempenho robusto, com o grupo classificadores individuais marcando um F1-Score médio de 94.29% ± 2.00 e o grupo comitê de classificadores registrando 99.16% ± 0.56. Os melhores classificadores foram o AD no grupo classificadores individuais e o XGB no grupo comitê de classificadores.
- **Análise das diferenças entre as bases de dados:**
  - **Cenários de baixo desempenho:**
    - No **Cenário 4 JSMA-FG**, ambos os conjunto de dados apresentaram os piores desempenhos gerais, indicando uma queda significativa no F1-Score quando o ataque FGSM foi combinado com a inserção de amostras do ataque JSMA na fase de treino. Para o *Power System Smart Grid Monitoring Power*, o grupo comitê de classificadores obteve um desempenho médio de 61.35%, enquanto no *Ereno IEC-61850 Intrusion Dataset*, o valor foi 99.60%. Isso evidencia uma diferença expressiva entre os dois conjunto de dados no mesmo cenário.

    - No **Cenário 3 FGSM**, o conjunto de dados *Power System Smart Grid Monitoring Power* apresentou uma queda acentuada, com um desempenho médio de 58.95% no grupo comitê de classificadores. Para o *Ereno IEC*, o desempenho foi significativamente melhor, com 99.65%. O comportamento entre os dois conjunto de dados também divergiu neste cenário.

- **Cenários de melhor desempenho:**
    - O **Cenário 4 JSMA-JSMA** destacou-se em ambos os conjunto de dados como o cenário de maior resiliência aos ataques adversários. Para o *Power System Smart Grid Monitoring Power*, o grupo comitê de classificadores alcançou 96.74%, enquanto no *Ereno IEC-61850 Intrusion Dataset*, esse valor foi ainda maior, atingindo 99.16%. Isso demonstra um comportamento consistente entre os dois conjunto de dados neste cenário.
    - No **Cenário 4 FG-FG**, os dois conjunto de dados também apresentaram alta resistência aos ataques. Para o *Power System Smart Grid Monitoring Power*, o grupo comitê de classificadores obteve 96.64%, enquanto no *Ereno IEC*, o resultado foi de 99.74%. Novamente, a consistência entre os conjunto de dados se mantém evidente.

- **Conclusão:**
    - Os cenários de baixo desempenho diferem significativamente entre os conjunto de dados. Enquanto o *Power System Smart Grid Monitoring Power* apresenta quedas mais acentuadas no desempenho, o *Ereno IEC-61850 Intrusion Dataset* demonstra maior resiliência, especialmente nos cenários envolvendo o ataque FGSM.
    - Nos cenários de melhor desempenho, há maior consistência entre os dois conjunto de dados, com ambos apresentando resultados elevados nos cenários **FG-FG** e **JSMA-JSMA**.

Os resultados indicam que a redução da quantidade de amostras de treino roubadas pelo atacante, no contexto do ataque FGSM, e a inserção de amostras adversárias geradas pelo ataque JSMA na fase de treinamento, combinadas com o ataque FGSM, foram os cenários que causaram a maior queda de desempenho dos classificadores em ambas as bases de dados analisadas. Além disso, os dados sugerem que a inserção de amostras adversárias no conjunto de treino proporciona melhorias no desempenho dos classificadores, com exceção da combinação que envolve a adição de amostras do ataque JSMA na fase de treino e a aplicação do ataque FGSM. Os melhores cenários foram aqueles em que as amostras adicionadas na fase de treino correspondiam ao ataque final, ou seja, **FG-FG** e **JSMA-JSMA**.

Os cenários de baixo desempenho diferiram significativamente entre as bases de dados, com o *Power System Smart Grid Monitoring Power* apresentando quedas mais acentuadas

no desempenho, especialmente no Cenário 4 JSMA-FG. Por outro lado, os cenários de melhor desempenho, como **FG-FG** e **JSMA-JSMA**, mostraram resultados consistentes e elevados em ambas as bases de dados. Isso evidencia que, embora os impactos variem, os cenários com maior resiliência foram similares entre as bases de dados analisadas.

As principais limitações deste trabalho incluem a ausência de análise do comportamento probabilístico dos algoritmos estudados, ou seja, a forma como esses algoritmos lidam com incertezas ou distribuições probabilísticas em suas tomadas de decisão. Além disso, não foram avaliadas variações no tamanho das amostras de teste e treino durante a execução dos métodos. Apesar dessas limitações, os resultados obtidos nos cenários analisados indicam que os algoritmos do grupo comitê de classificadores demonstraram maior eficácia na construção de IDSs aplicados a CPSs.

Capítulo 6

Conclusão

Este trabalho teve como objetivo analisar o impacto de ataques adversários em IDSs baseados em anomalias para CPSs. Os IDSs foram construídos utilizando classificadores do tipo classificadores individuais e comitê de classificadores, além de considerar duas bases de dados.

Os algoritmos de AM investigados neste estudo incluíram SVM, KNN e AD, representando os classificadores individuais, além de RF, GBM, AdaBoost e XGB, que compuseram o grupo comitê de classificadores. Para realizar os testes, foram utilizados dois tipos de ataques adversários: FGSM e JSMA. A avaliação dos classificadores foi conduzida utilizando o F1-Score como métrica de desempenho.

Este estudo apresentou uma análise abrangente sobre o impacto de ataques adversários em modelos de detecção de intrusão aplicados CPSs. A investigação destacou a vulnerabilidade de algoritmos de aprendizado de máquina, tanto classificadores individuais quanto comitê de classificadores, aos ataques FGSM e JSMA, evidenciando variações no impacto dependendo do classificador, da base de dados utilizada e da configuração dos parâmetros dos ataques. A pesquisa também avançou na avaliação da eficácia dos parâmetros *eps*, *theta* e *gama*, demonstrando que ajustes estratégicos nesses valores podem amplificar significativamente os efeitos adversos, contribuindo para um entendimento mais profundo das dinâmicas de ataque e defesa em CPSs.

Adicionalmente, a análise experimental incluiu cenários que consideraram a redução de amostras de treino roubadas e a inserção de amostras adversárias no treinamento, permitindo explorar a resiliência dos modelos. Os resultados indicaram que, embora a redução de amostras de treino nem sempre atenue as vulnerabilidades dos modelos, a inclusão de amostras adversárias pode fortalecer sua robustez contra futuros ataques. Essas contribuições forneceram informações valiosos para o aprimoramento de estratégias defensivas, apontando caminhos para o desenvolvimento de classificadores mais resilientes e destacando a importância de uma abordagem metodológica detalhada para proteger CPSs contra ataques adversários.

Os resultados demonstraram que os algoritmos do grupo comitê de classificadores

apresentaram desempenho superior em comparação aos classificadores individuais. No entanto, em algumas situações específicas, como nos ataques FGSM e em outros cenários descritos no Capítulo 5, os classificadores individuais apresentaram desempenho superior. Entre os melhores algoritmos de cada grupo, a AD destacou-se como o mais eficiente entre os classificadores individuais, enquanto o RF foi o algoritmo de maior destaque no grupo de comitê de classificadores.

6.1 Principais contribuições

As principais contribuições deste trabalho podem ser destacadas conforme a seguir:

1. Análise experimental do impacto em modelos de detecção de intrusão: Este estudo realizou uma análise experimental detalhada sobre o impacto de ataques adversários no desempenho de modelos de detecção de intrusão aplicados a sistemas ciberfísicos (CPSs). A investigação incluiu a avaliação de algoritmos de aprendizado de máquina (classificadores individuais e comitê de classificadores) sob diferentes cenários de ataques adversários, utilizando duas bases de dados distintas. Essa abordagem amplia o escopo da literatura existente, diferenciando este trabalho dos anteriores.

2. Identificação de vulnerabilidades nos algoritmos de aprendizado de máquina: Os resultados obtidos evidenciam que, de forma geral, os algoritmos de AM analisados apresentam vulnerabilidades frente aos ataques adversários FGSM e JSMA, demonstrando a necessidade de aprimoramento das estratégias de defesa em sistemas ciberfísicos.

3. Influência dos ataques adversários nos classificadores: O impacto dos ataques adversários variou significativamente conforme o classificador utilizado (classificadores individuais ou comitê de classificadores), a base de dados e a configuração dos parâmetros dos ataques. Essa observação fornece insights valiosos para a escolha de algoritmos mais robustos dependendo do cenário de aplicação.

4. Contribuição metodológica para avaliação de parâmetros de ataque: Este estudo investigou a influência de diferentes configurações dos parâmetros dos ataques (FGSM e JSMA) nos modelos analisados, evidenciando como ajustes nos parâmetros podem alterar significativamente a eficácia dos ataques e o desempenho dos classificadores.

5. Análise experimental do impacto da variação dos parâmetros *eps* do ataque FGSM e *theta* e *gama* do ataque JSMA no desempenho de modelos de detecção de intrusão em sistemas ciberfísicos. Este estudo destacou a relevância da configuração dos parâmetros *eps*, *theta* e *gama* na eficácia dos ataques adversários, identificando aqueles com maior influência nos resultados. A investigação detalhada de diferentes

variações e combinações desses parâmetros demonstrou que a otimização dos ajustes pode potencializar os efeitos adversos de maneira significativa. Essa análise fornece uma contribuição central ao campo, ao oferecer um entendimento aprofundado sobre como parâmetros ajustados de forma estratégica podem maximizar a eficácia de ataques adversariais. A análise evidencia que a configuração adequada dos parâmetros é crucial para amplificar a eficácia dos ataques adversários. Os insights gerados por este estudo podem orientar futuras pesquisas e estratégias defensivas, aprimorando tanto a compreensão sobre ataques quanto o desenvolvimento de modelos mais resilientes em sistemas ciberfísicos.

6. Análise experimental do comportamento de modelos de detecção de intrusão em sistemas ciberfísicos diante da diminuição na quantidade de amostras de treino roubadas. Os resultados indicaram que, na maioria dos cenários, a redução no número de amostras de treino roubadas não atenuou a vulnerabilidade dos modelos a ataques adversários. Com exceção do classificador SVM, a diminuição de amostras de treino roubadas no ataque FGSM resultou em uma redução no desempenho de todos os demais classificadores. Por outro lado, para o ataque JSMA, apenas os classificadores do grupo classificadores individuais foram afetados, com exceção do SVM. Esses resultados reforçam a importância de considerar as características específicas dos classificadores e ataques ao desenvolver estratégias de defesa para sistemas ciberfísicos.

7. Análise experimental do comportamento de modelos de detecção de intrusão em sistemas ciberfísicos diante da inserção de amostras adversárias no conjunto de treinamento. Os resultados indicaram que a inclusão dessas amostras pode contribuir para o aumento da robustez dos modelos em relação a ataques adversários.

8. Resultados que permitiram identificar pontos de melhoria e delinear estratégias voltadas para o aprimoramento dos classificadores, com o objetivo de aumentar sua resiliência frente a ataques adversários.

6.2 Contribuições em produção bibliográfica

Resultados parciais deste trabalho são descritos no artigo "*Robustness of Machine Learning-Based Intrusion Detection Systems Against Adversarial Attacks in Cyber-Physical Systems*", apresentado como artigo completo no ENIAC 2024 (*21st Encontro Nacional de Inteligência Artificial e Computacional*), que integra o BRACIS (*34th Brazilian Conference on Intelligent Systems*) – Qualis B4.

6.3 Trabalhos futuros

Os seguintes desdobramentos do presente estudo podem ser objetos de trabalhos futuros:

- ❏ Examinar o impacto de variações no tamanho dos conjuntos de teste e treino durante a execução do método, buscando compreender como essas alterações influenciam o desempenho dos classificadores.
- ❏ Avaliar o desempenho dos classificadores utilizando métricas adicionais, como revocação, precisão, taxa de falsos positivos e taxa de falsos negativos, para uma análise mais abrangente.
- ❏ Examinar o desempenho dos classificadores em diferentes fontes de dados, visando ampliar a generalização e a aplicabilidade dos resultados obtidos.
- ❏ Analisar o impacto de componentes probabilísticos nos algoritmos estudados, por meio da realização de múltiplos experimentos para calcular médias de desempenho, reduzindo os efeitos da aleatoriedade. Adicionalmente, identificar possíveis limitações e propor melhorias na forma como esses componentes influenciam os resultados dos modelos.
- ❏ Integrar os classificadores mais eficazes a um CPS, realizando testes em ambientes reais para coletar, armazenar e analisar dados, além de gerar alertas sobre atividades maliciosas provenientes de diversas fontes.
- ❏ Avaliar cenários alternativos aos considerados no presente estudo, explorando ambientes além do modelo *gray box*, para ampliar o espectro de aplicação do método.
- ❏ Realizar uma análise detalhada dos atributos presentes nos conjuntos de dados utilizados, com o objetivo de identificar quais características são mais suscetíveis ou resistentes a ataques adversários.
- ❏ Avaliar o impacto de ataques adversários no contexto de fluxos contínuos de dados (*data streams*)

Referências

ABUBAKAR, A.; PRANGGONO, B. Machine learning based intrusion detection system for software defined networks. In: IEEE. **2017 seventh international conference on emerging security technologies (EST)**. [S.l.], 2017. p. 138–143.

ALATWI, H. A.; ALDWEESH, A. Adversarial black-box attacks against network intrusion detection systems: A survey. In: IEEE. **2021 IEEE World AI IoT Congress (AIIoT)**. [S.l.], 2021. p. 0034–0040.

ALGULIYEV, R.; IMAMVERDIYEV, Y.; SUKHOSTAT, L. Cyber-physical systems and their security issues. **Computers in Industry**, Elsevier, v. 100, p. 212–223, 2018.

ALHAJJAR, E.; MAXWELL, P.; BASTIAN, N. Adversarial machine learning in network intrusion detection systems. **Expert Systems with Applications**, Elsevier, v. 186, p. 115782, 2021.

ALOTAIBI, A.; RASSAM, M. A. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. **Future Internet**, MDPI, v. 15, n. 2, p. 62, 2023.

ALSHAHRANI, E. et al. Adversarial attacks against supervised machine learning based network intrusion detection systems. **Plos one**, Public Library of Science San Francisco, CA USA, v. 17, n. 10, p. e0275971, 2022.

ANTHI, E. et al. Adversarial attacks on machine learning cybersecurity defences in industrial control systems. **Journal of Information Security and Applications**, Elsevier, v. 58, p. 102717, 2021.

AYUB MD AHSAN; JOHNSON, W. A. T. D. A. S. A. Model evasion attack on intrusion detection systems using adversarial machine learning. In: IEEE. **ANNUAL CONFERENCE ON INFORMATION SCIENCES AND SYSTEMS (CISS), 54.** [S.l.], 2020. p. 1–6.

BATISTA, G. E. d. A. P. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Tese (Doutorado em Ciência da Computação)) — Universidade de São Paulo, São Carlos, 2003.

BENETTI, T. **Segurança da Informação – Confidencialidade, Integridade e Disponibilidade (CID)**. 2015. Disponível em: <https://www.profissionaisti.com.br/

2015/07/seguranca-da-informacao-confidencialidade-integridade-e-disponibilidade-cid/ >.

BOLBOT VICTOR; THEOTOKATOS, G. B. L. M. B. E. V. D. Vulnerabilities and safety assurance methods in cyber-physical systems: A comprehensive review. **Reliability Engineering & System Safety**, Elsevier, v. 182, p. 179–193, 2019.

BRUNIALTI, L. et al. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: uma revisão sistemática. **Anais do XI Simpósio Brasileiro de Sistemas de Informação**, SBC, p. 203–210, 2015.

CÂMARA, S. D. M. **AUTENTICAÇÃO DE EVIDÊNCIAS DIGITAIS PARA SISTEMAS CIBERFÍSICOS DE RECURSOS LIMITADOS**. 123 p. Tese (Tese (Doutorado em Informática)) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, mar 2017.

CORONA I.; GIACINTO, G. R. F. Adversarial attacks against intrusion detection systems: taxonomy, solutions and open issues. **Information Sciences**, Elsevier, v. 239, p. 201–225, 2013.

CÔRTE, K. **Segurança da informação baseada no valor da informação e nos pilares tecnologia, pessoas e processos. 2014. 212 f., il**. 212 p. Tese (Tese (Doutorado em Ciência da Informação)) — Universidade de Brasília, Brasília, 2014.

EINY, S.; OZ, C.; NAVAEI, Y. D. The anomaly-and signature-based ids for network security using hybrid inference systems. **Mathematical Problems in Engineering**, Wiley Online Library, v. 2021, n. 1, p. 6639714, 2021.

ERDEMIR, E. et al. Adversarial robustness with non-uniform perturbations. **Advances in Neural Information Processing Systems**, v. 34, p. 19147–19159, 2021.

FIGUEROA, H.; WANG, Y.; GIAKOS, G. C. Adversarial attacks in industrial control cyber physical systems. In: IEEE. **2022 IEEE International Conference on Imaging Systems and Techniques (IST)**. [S.l.], 2022. p. 1–6.

GARCIA-TEODORO, P. et al. Anomaly-based network intrusion detection: Techniques, systems and challenges. **computers & security**, Elsevier, v. 28, n. 1-2, p. 18–28, 2009.

GIPIŠKIS, R. et al. The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems. **IEEE Systems Journal**, IEEE, v. 17, n. 4, p. 5327–5334, 2023.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. **arXiv preprint arXiv:1412.6572**, 2015.

GREER, C. et al. **Cyber-Physical Systems and Internet of Things**. [S.l.], 2019. Disponível em: <https://doi.org/10.6028/NIST.SP.1900-202>.

GUTTMAN, B.; ROBACK, E. A. **An introduction to computer security: the NIST handbook**. Diane Publishing, 1995. v. 800. Disponível em: <http: //dx.doi.org/10.6028/NIST.SP.800-12>.

HARIGUNA, T.; HANANTO, A. R. Improved intrusion detection system (ids) performance using machine learning: A comparative study of single classifier and ensemble learning. In: IEEE. **2022 IEEE Creative Communication and Innovative Technology (ICCIT)**. [S.l.], 2022. p. 1–7.

HELM, J. M. et al. Machine learning and artificial intelligence: definitions, applications, and future directions. **Current reviews in musculoskeletal medicine**, Springer, v. 13, p. 69–76, 2020.

HUANG, L. et al. Adversarial machine learning. In: **Proceedings of the 4th ACM workshop on Security and artificial intelligence**. [S.l.]: ACM, 2011. p. 43–58.

JAMAL, A. A. et al. A review on security analysis of cyber physical systems using machine learning. **Materials today: proceedings**, Elsevier, v. 80, p. 2302–2306, 2023.

JIA, Y. et al. Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. **International Journal of Critical Infrastructure Protection**, Elsevier, v. 34, p. 100452, 2021.

KARATAS, G.; DEMIR, O.; SAHINGOZ, O. K. Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset. **IEEE access**, IEEE, v. 8, p. 32150–32162, 2020.

KHAW, Y. M. et al. Evasive attacks against autoencoder-based cyberattack detection systems in power systems. **Energy and AI**, Elsevier, p. 100381, 2024.

KORONIOTIS, N. et al. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. **Future Generation Computer Systems**, v. 100, p. 779–796, 2019.

KRAVCHIK, M.; SHABTAI, A. Detecting cyber attacks in industrial control systems using convolutional neural networks. In: **Proceedings of the 2018 workshop on cyber-physical systems security and privacy**. [S.l.: s.n.], 2018. p. 72–83.

KUMAR, R.; SHARMA, D. Hyint: signature-anomaly intrusion detection system. In: IEEE. **2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)**. [S.l.], 2018. p. 1–7.

LI, J. et al. Adversarial attacks and defenses on cyber–physical systems: A survey. **IEEE Internet of Things Journal**, IEEE, v. 7, n. 6, p. 5103–5115, 2020.

LOPES, Y. et al. Smart grid e iec 61850: Novos desafios em redes e telecomunicaçoes para o sistema elétrico. **XXX Simpósio Brasileiro de Telecomunicações**, 2012.

MARTINS, N. et al. Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. **IEEE Access**, IEEE, v. 8, p. 35403–35419, 2020.

MOHANTY, H.; ROUDSARI, A. H.; LASHKARI, A. H. Robust stacking ensemble model for darknet traffic classification under adversarial settings. **Computers & Security**, Elsevier, v. 120, p. 102830, 2022.

MOUSTAFA, N.; SLAY, J. Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: **2015 Military Communications and Information Systems Conference (MilCIS)**. [S.l.: s.n.], 2015. p. 1–6.

NAKAMURA, E. T.; GEUS, P. L. de. **Segurança de redes em ambientes cooperativos**. [S.l.]: Novatec Editora, 2007.

NASCIMENTO, G.; CORREIA, M. Anomaly-based intrusion detection in software as a service. In: IEEE. **2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W)**. [S.l.], 2011. p. 19–24.

PACHECO, Y.; SUN, W. Adversarial machine learning: A comparative study on contemporary intrusion detection datasets. In: **ICISSP**. [S.l.: s.n.], 2021. p. 160–171.

PAPERNOT, N. et al. The limitations of deep learning in adversarial settings. In: IEEE. **2016 IEEE European Symposium on Security and Privacy (EuroS&P)**. [S.l.], 2016. p. 372–387.

PRANDEL, P. C. **Modelagem computacional para Avaliação da Age of Information (AoI) em sistemas ciberfísicos**. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2022.

QUINCOZES, S. E. et al. Ereno: An extensible tool for generating realistic iec-61850 intrusion detection datasets. In: SBC. **Anais Estendidos do XXII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais**. [S.l.], 2022. p. 1–8.

_____. ERENO: A framework for generating realistic IEC-61850 intrusion detection datasets for smart grids. **IEEE Transactions on Dependable and Secure Computing**, v. 21, n. 4, p. 3851–3865, 2024.

ROCHA, M. S. et al. Supervised machine learning and detection of unknown attacks: an empirical evaluation. In: SPRINGER. **International Conference on Advanced Information Networking and Applications**. [S.l.], 2023. p. 379–391.

ROUZBAHANI, H. M. et al. Anomaly detection in cyber-physical systems using machine learning. **Handbook of big data privacy**, Springer, p. 219–235, 2020.

SAHANI, N. et al. Machine learning-based intrusion detection for smart grid computing: A survey. **ACM Transactions on Cyber-Physical Systems**, ACM New York, NY, v. 7, n. 2, p. 1–31, 2023.

SARANYA, T. et al. Performance analysis of machine learning algorithms in intrusion detection system: A review. **Procedia Computer Science**, Elsevier, v. 171, p. 1251–1260, 2020.

SILVA, G. H. E. da; MIANI, R. S.; ZARPELAO, B. B. Investigando o impacto de amostras adversárias na detecção de intrusões em um sistema ciberfísico. In: SBC. **Anais do XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**. [S.l.], 2023. p. 281–294.

SILVA, J. Inteligência artificial e aprendizado de máquina. **Estudos Avançados**, Scielo, v. 37, n. 107, p. 45–67, 2023.

STOLFO, S. J. et al. Cost-based modeling for fraud and intrusion detection: Results from the jam project. In: **DARPA Information Survivability Conference and Exposition**. [S.l.: s.n.], 2000. v. 2, p. 130–144.

STOUFFER, K. et al. Guide to industrial control systems (ics) security. **NIST special publication**, v. 800, n. 82, p. 16–16, 2011.

TAVALLAEE, M. et al. A detailed analysis of the kdd cup 99 data set. In: **Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)**. [S.l.: s.n.], 2009. p. 1–6.

UÇAR, M. K. et al. The effect of training and testing process on machine learning in biomedical datasets. **Mathematical Problems in Engineering**, Hindawi, v. 2020, 2020.

UTIMURA, L. N. **Aplicação em tempo real de técnicas de aprendizado de máquina no Snort IDS**. Dissertação (Mestrado) — Universidade Estadual Paulista (Unesp), 2020.

VALEEV, N. Bachelor's thesis, **Systematic Literature Review of the Adversarial Attacks on AI in Cyber-Physical Systems**. 2022. Disponível em: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1671279>.

WANG, S. et al. Evasion attack and defense on machine learning models in cyber-physical systems: A survey. **IEEE Communications Surveys & Tutorials**, IEEE, 2023.

WANG, Z. Deep learning-based intrusion detection with adversaries. **IEEE Access**, IEEE, v. 6, p. 38367–38384, 2018.

WOLDEYOHANNES, H. D. Review on "adversarial robustness toolbox (art) v1. 5. x.": Art attacks against supervised learning algorithms case study. Università Ca'Foscari Venezia, 2021.

YANG, K. et al. Adversarial examples against the deep learning based network intrusion detection systems. In: IEEE. **MILCOM 2018-2018 ieee military communications conference (MILCOM)**. [S.l.], 2018. p. 559–564.

ZHANG, J. et al. Deep learning based attack detection for cyber-physical system cybersecurity: A survey. **IEEE/CAA Journal of Automatica Sinica**, IEEE, v. 9, n. 3, p. 377–391, 2021.