TOWARDS PHYSICS-BASED INDUSTRIAL MALWARE DETECTION WITH PERFORMANCE AND INTEGRITY ANALYSIS ENHANCED VIA MACHINE AND DEEP LEARNING

by

HUSSAM ABED-ALRAHMAN EID AL-TARAZI

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE AND INFORMATICS

2025

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Julian Rrushi, Ph.D., Chair
Hua Ming Ph.D.
Ilias Cholis, Ph.D.
Amine Barrak, Ph.D.

© Copyright by Hussam Abed-AlRahman Eid Al-Tarazi, 2025
All rights reserved

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Julian Rrushi, and the rest of the committee members namely Dr. Ilias Cholis, Dr. Amine Barrak and Dr. Hua Ming. Dr. Julian has been a wonderful mentor, coach, and motivator throughout my PhD journey. Julian's research and scientific expertise is exceptional and invaluable, and this research would not have been possible without his support and guidance.

I also would like to thank my family: my passed mother and father and my brothers Dr. Eid and Nizar and my sisters for supporting me financially and morally throughout my journey.

Hussam Tarazi

ABSTRACT

TOWARDS PHYSICS-BASED INDUSTRIAL MALWARE DETECTION WITH PERFORMANCE AND INTEGRITY ANALYSIS ENHANCED VIA MACHINE AND DEEP LEARNING

by

Hussam Abed-AlRahman Eid Al-Tarazi

Advisor: Julian Rrushi, Ph.D

The security of power system protection infrastructure poses emerging challenges, particularly due to the implications of threats that exploit the physical behavior of power transformers. While prior research has made significant progress in detecting malware within industrial control systems (ICS), there remains a pressing need for more robust and adaptive methodologies capable of addressing evolving cyber threats and advanced intrusive techniques.

This study emphasizes the critical role of protection relay algorithms specifically differential and harmonic restraint algorithms in safeguarding power transformers from electrical faults. These algorithms, however, are susceptible to malicious interference; they can be disabled, altered, or suppressed by malware, thereby compromising the reliability and safety of the system.

We begin by presenting an overview of a typical electrical power substation architecture and the core functionalities of protection relay algorithms. Furthermore, we explore preliminary insights into malware strategies that involve physics-based data

manipulation, potentially leading to transformer maloperation, large-scale blackouts, and permanent equipment damage.

To address these vulnerabilities, we propose the design of a hybrid machine learning and deep learning framework aimed at detecting protection anomalies through analysis of computational resource metrics, including memory usage and execution time, associated with relay algorithm execution. The model was developed using Python and evaluated through simulation of differential and harmonic protection logic. Synthetic datasets were generated using a generative adversarial network (GAN) implemented on a virtualized environment to facilitate comprehensive model training and testing.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES – Continued

LIST OF ABBREVIATIONS

CPS Cyber-physical system

IoT Internet-of-things

AI artificial intelligence

ICS Industrial control system

DDoS Distributed denial of service

OT Operational technology

SCADA Supervisory control and data acquisition

CIA triangle (Confidentiality, integrity, and availability)

RBA Role-based access

PLC Programmable Logic Controllers

HMI Human-Machine Interfaces

IDS Intrusion detection system

MFA policies and multi-factor authentication

AC Alternating current

DC Direction current

MVA Megavolt ampere

ADC Analog to digital converter

RAM Random access memory

CT Circuit Transformer

KCL Kirchhoff's circuit law

FFT Fast Fourier transform

LIST OF ABBREVIATIONS - Continued

SIS Safety instrumented system

I/O Input/output

RMS Root mean square

C2 Command & control

VPN Virtual private network

PNN Probabilistic Neural Network

TSV Trusted Safety Verifier

GANS Generative adversarial networks

CGANS Conditional generative adversarial networks

CHAPTER ONE: INTRODUCTION

Cyber-Physical Systems (CPS) are intelligent systems integrating computational and physical components, functioning in close interaction to monitor real-world scenarios. These systems exhibit high complexity across various spatial and temporal scales, with extensive networked communications that merge computational and physical elements [5]. Embedded computers and networks oversee and regulate physical processes, typically incorporating feedback loops where physical actions influence computations and vice versa. CPS systems are ubiquitous across industries where digital systems are integrated into traditional analog domains. The Internet of Things (IoT) exemplifies this trend, embedding micro-controllers in everyday objects like smartphones, lightbulbs, and air conditioners. In industrial contexts, this digital control and monitoring is termed Industrial Control Systems (ICS).

Within the electric power sector, ICS systems have evolved to perform diverse tasks within energy control centers. When applied to power grids, these systems create what is known as a smart grid. Research into using digital computers for power system protection tasks began in the mid-1960s, driven by advancements in process-control computers [7]. The adoption of microcontrollers, replacing electromechanical and solid-state relays, offers several advantages including enhanced data for analysis, algorithm refinement for improved performance, and cost efficiencies [8].

These systems based on microcontrollers have evolved into cyber-physical systems, enabling enhanced monitoring and control capabilities over longer distances. However,

the adoption of microcontroller-based protection methods also introduces vulnerabilities, potentially allowing cyber threat actors to exploit unprotected systems. This poses risks to smart grids, enabling unauthorized access to sensitive data or control of physical processes in critical industrial facilities such as distribution substations.

To counter these cyber threats, robust defense strategies must be developed to safeguard the power industry, ensuring the continuous delivery of reliable, secure, and affordable electricity to consumers. As cyber threat actors grow more sophisticated, many exploit specialized knowledge of customized CPS architectures and leverage physical principles to launch attacks on smart grids [1]. Our focus is on exploring methods through which threat actors could use their understanding of physics-based principles to compromise the protection algorithms used in power transformers. Additionally, we will perform performance and integrity analysis for the protection algorithms to understand how these algorithms can resist cyber-attacks, more specifically, physics-based malware designed to alter or affect the accuracy and integrity of these algorithms output, and the underlying resources needed for running them.

Initially, we conduct threat modeling of the cyber-physical ecosystem to identify potential threats and attack vectors that threat actors might employ. Subsequently, we develop a software simulation environment that accurately replicates the operational physics of a power transformer, including the specific physics-based protection algorithms.

Following this, we develop adversarial code informed by physics principles to simulate 2 different potential attacks on these protection algorithms. We rigorously test this

adversarial code to assess its effectiveness in circumventing the implemented protective measures.

Cybersecurity challenges in smart grids

Smart grids represent the next generation of electrical power systems, integrating advanced information and communication technologies to enhance the efficiency, reliability, and sustainability of electricity services. However, the integration of these technologies introduces significant cybersecurity challenges. The increased connectivity and digitalization render smart grids more susceptible to cyberattacks, which can lead to severe consequences such as physical damage, service interruptions, and even threats to human safety [1]. In the context of smart grids, the well-known CIA (Confidentiality, Integrity, Availability) triangle is often inverted. Availability is paramount because electricity structures ought to perform continuously to make sure of solid energy delivery. Integrity is critical as any inaccuracies in measurements or control instructions can bring maloperations [2]. The specific traits of smart grids, together with their allotted nature and the combination of renewable energy assets, similarly complicate the cybersecurity landscape [3].

The extended connectivity in smart grids introduces new vulnerabilities. Traditional energy grids had been quite remote structures with restricted external connections. In comparison, smart grids rely on massive conversation networks to facilitate real-time information alternate and remote-control abilities. These networks consist of various interconnected devices, including smart meters, sensors, and automation management equipment, which may be open access points for cyberattacks [4]. Cyberattacks on smart grids can take diverse schemes, along with malware infections, denial of service (DoS) attacks, and phishing maneuvers. These attacks can disrupt the everyday operation of

smart grids, leading to energy outages, harm to physical devices, and compromised information integrity [5]. For instance, malware can infiltrate the grid's control systems and alter critical settings, causing cascading failures across the network [6].

To expand more about the CIA triangle (Confidentiality, integrity, and availability) in ICS's figure 1, they are the three pillars of cybersecurity, commonly referred to as the CIA triangle. In the context of smart grids, these pillars are essential to maintaining the security and reliability of the system. However, the prioritization of these elements may differ from traditional IT environments. In smart grids, availability is the highest priority. The continuous operation of power systems is crucial to ensure a stable electricity supply. Any deterioration in the availability of resources can have serious consequences, including power outages, economic losses, and public safety risks [7]. Authenticity is important in smart networks because accurate and reliable information is important for proper and valid control and assessment. Incorrect feedback or unauthorized changes can open doors to attackers, including incorrect modifications or false alarms. While confidentiality is important, it is often considered a low priority as compared to the provision and integrity of smart grids.

Figure 1: CIA Triangle

The variety of cybersecurity threats facing smart grids is continually evolving. These threats can originate from various belongings, which include insider threats, state actors, and cyber-criminal groups. Understanding the nature of these threats is vital for developing effective countermeasures. Insiders with privileged access to critical systems pose a massive hazard to smart grid protection. These people may intentionally or accidentally compromise devices via disclosing sensitive information or misconfiguration [1]. State-actors often have extensive resources and expertise, making them able to launch zero day cyberattacks on smart grids. These actors may additionally target the power infrastructure to benefit geopolitical desires, inclusive of disrupting the electricity supply of an adversary [2]. Cyber-criminal groups or companies might also target smart grids for financial gain. For instance, ransomware attacks can encrypt data and ask for a ransom for its decryption. Such attacks can disrupt grid operations and motive monetary losses [3].

To address those threats, various cybersecurity measures can be carried out. These measures include Intrusion Detection Systems (IDS), that may display network traffic for signs and symptoms of malicious activity and alert operators to capability threats.

Advanced IDS use device gaining knowledge of algorithms to find anomalies and improve threat detection accuracy [4]. Encrypting communication channels and sensitive information can save us from unauthorized access and ensure the confidentiality and integrity of information. Implementing strong policies can restrict the ability of intruders to access essential structures and information. Role-based access (RBA) policies and multi-factor authentication (MFA) can minimize the risk of unauthorized access to systems; they are considered powerful techniques for boosting access rights to control systems [6]. Additionally, keeping software applications and firmware updated is vital for shielding against new and unpatched vulnerabilities. Developing and often updating incident response plans can help organizations rapidly and effectively respond to cybersecurity incidents. These plans need to include strategies for identifying, containing, and mitigating cyberattacks [8].

As smart grids continue to adapt, ongoing studies are wanted to cope with rising cybersecurity demanding situations. Future studies guidelines include developing enhanced cyber detection algorithms that use machine learning and artificial intelligence to come across and reply to sophisticated cyberattacks [2]. Building stable communication systems which can resist and recover from cyberattacks is another priority. This includes the use of redundancy and fail-safe mechanisms to support systems' continuity and availability [3]. Combining cybersecurity measures with physical security practices to provide comprehensive protection against cyber-physical threats is also an important research direction. This approach recognizes the interaction between cyber-physical systems in smart grids [4]. Finally, growing and enforcing cybersecurity

guidelines and policies that mandate pleasant practices and standards for smart grid protection is required. Collaboration between authorities' businesses, industry stakeholders, and academia is deemed important for developing a steady and resilient electricity infrastructure [5].

Industrial Control Systems Malware

Industrial control systems (ICSs) are critical components of smart grids, answerable for monitoring and controlling physical assets. These systems are getting more and more targeted by malware, which can motive significant damage via the use of switching rules, disrupting performance, and inflicting massive failures [16]. ICS typically include SCADA (monitoring control and data acquisition) architecture, DCS (distributed control system), PLC (Programmable Logic Controllers), HMI (Human-Machine Interfaces), and SIS (Safety Instrumented Systems) [17]. As we mentioned earlier, the integration of these systems into the smart grid system introduces cybersecurity vulnerabilities that can be exploited by malicious actors.

Protection relays, vital for protective energy systems, are especially prone. These digital devices, included in the SCADA architecture, may be compromised by malware, leading to altered settings and disrupted power supply [18]. The SCADA architecture for power systems (Figure 1) generally consists of three levels: Station Level, Bay Level, and Process Level. At the Station Level, the structure includes the Human-Machine Interfaces (HMI) for actual-time tracking, engineering stations. The Bay Level homes digital relays, and Intelligent Electronic Devices (IEDs) for safety, management, and automation. The Process Level includes merging units, circuit breakers, and power transformers [19].

Each of these levels has specific vulnerabilities that can be exploited by attackers to compromise the ICS.

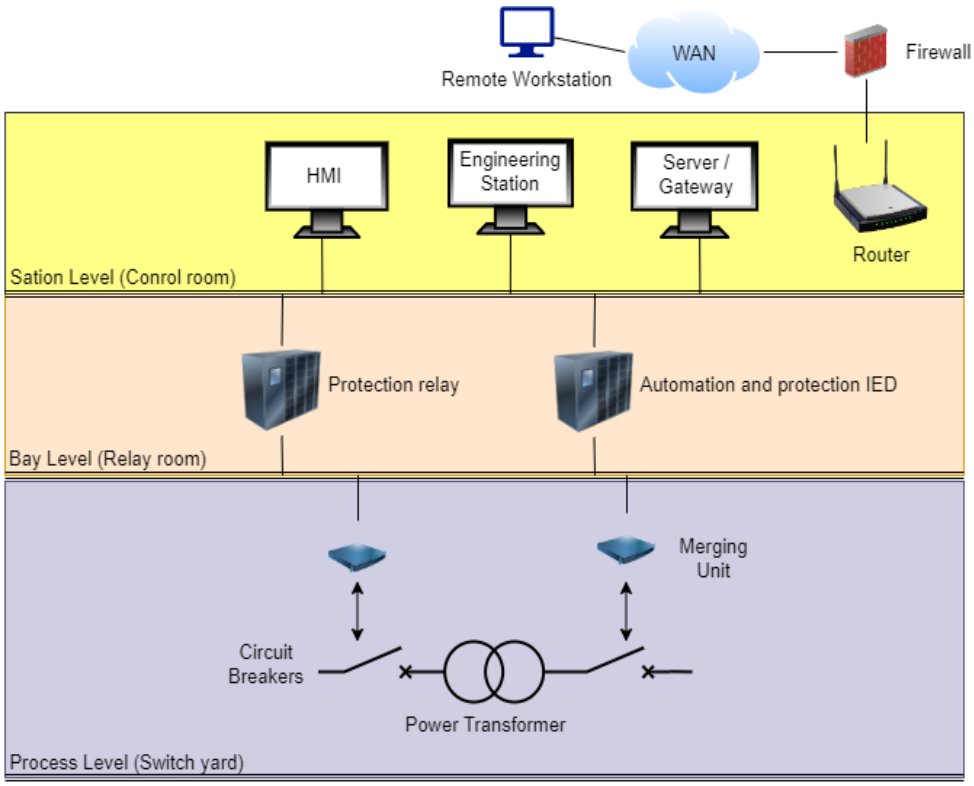


Figure 2: SCADA power substation

The station level is critical to the overall functionality of the SCADA system, as it provides an interface for human operators to monitor and control the system in real time, HMIs at this level can be targeted by malware to provide false data or compromised critical controls, leading to wrong decisions [17]. It is important to maintain the integrity and availability of the ICS to ensure the security of these resources.

The Bay Level, wherein digital relays and industrial automation equipment's are placed, is chargeable for the actual management and automation of the electricity device. Digital relays are mainly critical as they provide protection against faults and abnormal

situations. However, those relays can be compromised via malware that alters their settings, inflicting them to trip incorrectly or fail to trip whilst needed [20]. This can lead to extensive power outages and damage to power transformers. Industrial devices and IEDs at this level are also susceptible to attacks which could disrupt their ordinary operation, leading to significant operational disruptions.

At the Process Level, merging units and circuit breakers play a critical role if the physical signals and measurements of the power transformer were manipulated. These components are liable for the real switching operations that manage the flow of electricity. Malware focused on those equipment's can cause them to operate incorrectly, leading to physical damage to devices and extended energy outages [21]. Ensuring the security of the Process Level is essential to maintaining the overall reliability and safety of the power system.

One prominent example of malware focused on ICSs is the Stuxnet computer virus, which focused on PLCs used in uranium enrichment centers. Stuxnet verified the capacity for malware to reason physical damage with the aid of manipulating settings and disrupting ordinary operations [17]. This assault highlighted the need for sturdy cybersecurity measures in ICSs to protect them from incidents. The sophistication of Stuxnet and its potential to stay away from detection for a prolonged period underscores the challenges in defending ICSs from advanced persistent threats.

The effect of malware on ICSs can be profound, affecting not just the operation of the electricity grid but additionally safety and security. For instance, in 2015, a cyberattack at the Ukrainian smart grid led to widespread electricity outages, affecting loads of lots of customers [19]. This incident confirmed the potential for malware to cause significant

disruptions and highlighted the vulnerabilities in ICSs that need to be addressed. The attackers had been able to gain access to the SCADA structures and manipulate control settings, causing the intentional shutdown of numerous substations.

To address the evolving risk of malware targeting ICSs, numerous cybersecurity solutions have been proposed. One approach is to activate advanced intrusion detection systems that can analyze network traffic for signs of malicious activity. These IDS's use machine learning algorithms to hit upon anomalies and become privy to cyber threats in real time [21]. By constantly checking network traffic and access logs, IDS can provide early warning of cyber-attacks and counter their ability to cause widespread harm.

Another important factor of securing ICSs is the implementation of strong access polices. Role-based access can totally prioritize privilege to control systems (RBAC) and multi-factor authentication (MFA) can help restrict access to essential systems and make certain that criminal personnel could not make changes or manipulate settings [18]. By restricting proper access to critical components, companies can reduce insider threats and stop unauthorized users from compromising the ICS. Regular audits and reviews of access policies, and authenticated users are essential to retaining strong access policies in the surrounding environment.

In addition to those technical measures, organizations should additionally keep recognition on developing whole cybersecurity awareness and strategies. Regular training and consciousness programs for personnel can assist in making sure that they recognize the risks associated with malware and the significance of following effective practices for cybersecurity [17]. Incident response plans need to be implemented and frequently

updated to make sure that entities are prepared to respond on time and correctly to cyberattacks. These plans must include tactics for detecting, containing, and mitigating the impact of malware on ICSs.

The integration of advanced technologies, which includes artificial intelligence (AI) into ICSs offers new possibilities for enhancing cybersecurity. AI may be used to develop predictive models that become aware of vulnerabilities and suggest proactive measures to cope with them [16]. Machines learning algorithms can examine datasets to perceive patterns and anomalies that can uncover a cyberattack. By leveraging these technologies, agencies can improve their capacity to stumble on and reply to threats in actual times.

Despite those improvements, there are still some challenges in securing ICSs from malware. One of the main demanding situations is the complexity of those structures and the required testability and dependency for protection measures. Implementing robust cybersecurity measures can often warfare with the need for real time malware analysis, making it difficult to attain a suitable balance [19]. Additionally, the evolving nature of cyberthreats approaching ICSs infers a continuous attack surface and coping with new vulnerabilities and threat actors might still be a challenge.

Physics-based cyber attacks

Physics-based attacks exploit the physical properties and system dynamics of electric power systems, making them uniquely vulnerable and important to the cybersecurity of these systems. Unlike traditional cyberattacks, physics-based attacks are real. These attacks can cause physical effects, such as causing disruptions in power transformer protection algorithms or triggering false tripping of relays [10]. These attacks are based

on the physical attributes of cyber-physical systems (CPS) and can leverage processes and control logic, making their detection and mitigation using traditional cybersecurity measures particularly difficult.

One of the key aspects in power transformer protection is the differential protection algorithm, which compares input and output transformer currents. This algorithm is important for rapidly detecting and keeping transformer faults apart, thereby preventing faults [22]. The differential algorithm set of laws operates on the precept that the difference between the input and output currents must be zero underneath regular situations. Any noticeable deviation from this predicted threshold indicates a fault. However, if this set of laws is compromised, it may fail to locate faults, leading to unsafe results, which include prolonged transformer damage and accelerated risk of cascading failures throughout the grid.

Another important aspect of power transformer protection is the harmonic protection algorithm, which is designed to detect and isolate harmonic currents. Harmonics are current waveforms that are many times the fundamental frequency, and can cause serious problems such as overheating, increased losses, and reduced efficiency. Ensuring integrity of these algorithms is important for efficient and reliable systems performance [12]. Harmonic currents also can cause equipment damage and, if not maintained, can lead to reduced efficiency, which underscores the importance of robust safety equipment.

The complexity and class of physics-based threats have been highlighted in numerous studies. For example, [14] advise a formal approach to reading physics-based cyber incidents in cyber-physical systems. Their findings emphasize the need for a

comprehensive framework of each of the cyber elements of CPS to correctly counter those threats. They argue that conventional cybersecurity measures, which target traditional IT infrastructure, are inadequate for protecting CPSs from physics-based threats, which exploit the gap among predictive and real time methods.

Similarly, [11] offer a survey of physics-based comprehensive attack detection and propose new directions for studies. They highlight the constraints of zero-day detection mechanisms and recommend the integration of irregular physical data patterns targeting industrial equipment's to enhance detection of such threats. By combining data-driven techniques with physics-based techniques, it is viable to have a robust and resilient detection framework that may filter out anomalies' indicative of an attack.

Authors in [12] explore a hybrid method that combines physics-based intrusion detection and information gathering techniques for mitigating the impact of cyberattacks on automation and control mechanisms in electricity structures. Their work demonstrates that leveraging models and analytical records can improve the resilience of CPSs against sophisticated attacks. This hybrid method allows for the detection of diffused anomalies that are probably neglected through systems, thereby improving the general security posture of the control system.

The importance of information assurance of the physical dynamics of a CPS in context of cybersecurity is in addition highlighted by [15], who check out physics-conscious mimicry attacks. These attacks take advantage of the physical attributes of equipment's to mimic regular operations at the same time as executing malicious operations. Such threats are particularly hard to encounter because they're able to run stealthily. By being

aware of the physical dynamics of devices, attackers can craft more sophisticated and stealthy threats which are harder to understand and mitigate.

In addition to the specific algorithms and approaches used for protection, the general architecture of CPS's plays an important role in their resilience to physics-based threats. The author in [11] suggest that adopting a unified scheme of protection, which integrates cyber-physical threats into the format and operation of CPS, can drastically enhance their robustness. This includes implementing redundant structures, failover mechanisms, and real-time monitoring systems that could detect and respond to intrusions.

The research work by [13] on hybrid mitigation strategies highlights the potential of mixing multiple techniques to gather a greater comprehensive security scheme. By integrating physics-based detection with machine and deep learning, it's viable to create adaptive systems which could examine and learn from past incidents and improve their detection and reaction abilities over time. This adaptive approach is especially essential in the context of evolving threats, where attackers continuously adopt new techniques to skip existing defenses.

Despite the advancements in detection and mitigation strategies, there are though enormous worrying situations in addressing physics-based attacks. One of the main demanding situations is the complexity of systems and component communication. This requires a deep knowledge of each domain and the ability to combine disparate features into a unified framework. [14] spotlights the need for interdisciplinary studies that brings together wide knowledge base from cybersecurity, control and power engineering to deal with the prospect threats correctly.

The evolving nature of cyber threats additionally necessitates continuous updates to protection frameworks. As new vulnerabilities and attack vectors are developed, it is vital to make sure that protection algorithms and detection mechanisms are frequently up to date to address these emerging threats. [15] emphasizes the importance of proactive safety and protection functions, which consist of everyday security audits, penetration tests, and the improvement of incident response plans that encompass situations for physics-based threats.

Problem statement

Protection relays serve as a foundational component in safeguarding power transformers, which are among the most critical assets in modern power systems. In various digital substation architectures, power transformers are the primary elements requiring protection. Circuit breakers are actuated by protection relays based on decisions generated by embedded protection algorithms. These algorithms are responsible for detecting faults and ensuring safe system operation by initiating appropriate protective actions.

As previously discussed, the physics-based logic within protection algorithms alongside execution time and memory usage forms the core of our concern when addressing anomaly detection and cyber-intrusion scenarios. It is imperative that the decisions or tripping commands produced by these algorithms are both valid and reliable. A failure in the algorithmic logic may compromise the entire protection mechanism, potentially leading to severe system instability and even catastrophic outcomes.

Several factors can undermine the proper functioning of protection algorithms, including various types of malware and targeted cyber-attacks. In electrical power systems, scenarios that may trigger circuit breakers and isolate transformers include the following:

1- Short circuits and overcurrent conditions, as well as inrush currents induced by harmonic distortion.
2- Transient states occurring during transformer energization, which can produce high inrush currents that may falsely trigger relay actions.
3- Cyber-attacks, such as false data injection that result in unauthorized tripping commands.
4- Noise or timing errors that cause relays to issue incorrect decisions, deviating from expected logic.
5- Cooling system failures, which may initiate tripping; however, such cases fall under thermal protection systems and are outside the scope of this research.

These scenarios pose significant risks, potentially leading to unintentional outages and even large-scale blackouts. Conversely, if a protection relay fails to respond to genuine fault conditions due to suppression by malicious code or incorrect logic the integrity and stability of power transformers may be severely compromised. This condition, known as suppressive behavior [22], poses long-term risks to both equipment and system stability.

The main research focus of this study is the detection of intrusive behaviors and physics-based malware specifically targeting protection algorithms. Existing literature highlights various methods through which protection systems can be deceived or disrupted. A specific concerning class of threats called physics-aware malware is engineered to alter

the logical behavior of industrial protection mechanisms to inflict physical damage, asset loss, and potentially threaten human safety. These malware types are notable for their ability to interfere directly with the execution of physical processes.

Industrial Control Systems (ICSs) are well-documented as high-value targets for cyber-attacks. Among the most prevalent attack types reported in the literature are false data injection (FDI) and denial of service (DoS) attacks [23]. However, these classifications encompass a wide range of attack strategies, including:

1- Network-level attacks: packet spoofing, DoS, false packet injection, and side-channel exploits.
2- Application-level attacks: broken authentication, malware outbreaks, false data injection, command injection, and ransomware.
3- Firmware-level attacks: unauthorized firmware modification, BIOS/UEFI manipulation, code reuse, and supply chain exploitation.

Prior research has investigated the growing prevalence of physics-based cyber-attacks on power transformers. For example, one study [24] detailed how an adversary could manipulate normal operating parameters, causing overloading and premature aging by inducing faults or tripping transformers under peak load conditions.

To date, eight major ICS-targeting malware families have been documented: Stuxnet, Havex, BlackEnergy, BlackEnergy2, CrashOverride, Trisis, PipeDream, and Industroyer [25]. Each of these malware families was developed with the intent to disrupt or damage the physical processes managed by ICS environments.

More recently, the emergence of Incontroller (also known as Pipedream) has been particularly alarming [26]. Designed to compromise energy sector infrastructure, this malware can target a broad array of industrial systems. Additionally, the identification of Industroyer2 which was deployed in a cyber-attack against a Ukrainian energy provider further underscores the sophistication and intent of modern adversaries. This malware aimed to cause tangible physical destruction through direct manipulation of protection mechanisms [9].

A typical example of such an attack might involve compromising an industrial computer that hosts protection algorithms. During a fault event, the compromised system fails to initiate breaker tripping, thereby allowing continued operation under hazardous conditions. This suppressive behavior has been observed in malware variants such as Industroyer [26] and Triton [24], both of which were engineered to bypass safety mechanisms and facilitate physical damage.

Research objectives

The primary objectives of this research are:

- Study and research the literature for state-of-the-art physics-based malware detection techniques
- Collect a dataset of normal and malicious relay operating conditions for the purpose of training and evaluating the intended malware detection scheme
- Conduct research based on a simulated environment where protection algorithms are emulated for different protection scenarios

We focus on advancing the current literature with a model for detecting and predicting physics-based malware and intrusive behavior in protection relays.

We will model and emulate different cyberattacks against protection algorithms before and during runtime and after activating the proposed malware threat which aims to disable or bypass the algorithms. We have three different attack variations to be emulated:

The freeze attack, the mask attack, and memory consumption attack on the running host system.

Contribution

Our view of the proposed model to detect malicious physics data is based on deep exploration of the inner logic of protection algorithms for a power transformer protection relay; more specifically, differential and harmonic protection algorithms. Few research papers addressed the security and integrity of protection algorithms and profiling its execution with correspondence to resource metrics, such as processor execution time and memory allocation and consumption. Machine and deep learning models play an important role in detecting and predicting anomalies and deviations in cyber physical systems and malware targeting such systems usually hides their presence and run stealth. Machine learning can assist with uncovering possible deviations in a protection relay by recognizing anomalous patterns during execution.

This work achieves two milestones. First, advancing literature with a model to detect physics-based cyber-attacks and uncovering the ways such attacks can target differential and harmonic protection algorithms by emulating resource intensive malware and a

cyber-attack affecting execution time. Secondly, it provides a method to assess protection algorithms' security and integrity based on different factors, such as generative ai and deviations detection through pattern recognition and prediction. Our proposed work is among the first approaches that provide a realistic assessment of protection algorithms that consume resources, such as memory and execution time, together with a sample of malware and cyber-attacks on power grid equipment. Our work emphasizes the evaluation of how the detection algorithm will react and the possibility of activating leveraged attacks to model better defense from current and future malware.

We emulated malware and two cyber-attacks that can be activated without the risk of damaging physical equipment. This allows for an unlimited number of experiments that can be customized and modified to test any factors of interest, cyber or physical, without concerns.

These efforts provide the cybersecurity community with valuable insights into how cyber-attacks can manipulate physics data to disrupt and affect microcontroller-protected power transformers. By implementing these attacks in a software environment, researchers can explore detailed scenarios crucial for designing effective defense strategies.

Organization of the dissertation

The subsequent chapters of this thesis are structured as follows: Chapter 1 provides a background on power transformers and related cyber security challenges as well as physics-based cyber threats. In chapter 2 we outlined the top collected literature review related research in the field, then we provided a background about protection algorithms

and the consequences of failures in power transformers, then we went through machine and deep learning models for detection and challenges facing such approaches.

Chapter 3 outlines the threat model for our emulation, detailing assumptions and capabilities of the simulated threat actors. Chapter 4 outlines the main aspects of the experiment. In Chapter 5, we thoroughly explored the detection approach with machine and deep learning. Chapter 6 discusses how we evaluated and tested the approach, and Chapter 7 concludes with final outcomes and remarks.

CHAPTER TWO: BACKGROUND AND LITERATURE REVIEW

Power transformers in a typical SCADA architecture

Power transformers are devices designed to convert current or voltage to a multiplied or a reduced value. This output value is a current or voltage that is higher or lower than the input, and a transformer is identified as being either "step-up" or "step-down", respectively, this is very useful to power companies that need to transmit higher voltage electrical power over long-distance lines to consumers at relatively low voltage. This change in value is determined by the ratio of the number of turns of the wires that are wrapped around the core of a conductive material, often steel or iron. This core provides a magnetic path to connect the flux created by induction [27]. The input current is referred to as the primary side, and the output current is the secondary side, these two coils are isolated from each other so no current can transfer directly.

According to Faraday's Law of Induction [28], when an AC current is passed through one coil, electrical energy in the other is induced by the magnetic flux that is created in the shared core. This flux creates an electromotive force on the secondary side. This phenomenon is known as electromagnetic induction and allows for a transference of electrical power between two circuits that are not physically connected to each other and the ability to step the current or voltage either up or down as needed for the situation [29]. This is the reason why AC is required by the electric transmission system: only a constantly changing current will create an electric field and thereby induce current in nearby conductors. By Faraday's Law of Induction, the primary side voltage can be written as: $V_p = -N_p \Delta\Phi / \Delta t$, and similarly the secondary side voltage as $V_s = -N_s \Delta\Phi /$

Δt, where N is the number of turns in the coil of wire on the primary or secondary side, and $\Delta\Phi / \Delta t$ is the derivative of the magnetic flux Φ through one turn of the winding over time. Combining these equations, we have Vs / Vp = Ns / Np, or that the ratio of voltages equals that of the ratio in the number of turns. The law of conservation of energy requires that the energy (and power) that enters the system be equal to the energy that exits, so we can write W = Vs * Is = Vp * Ip. By combining terms, we conclude that Vs / Vp = Ns / Np = Ip / Is. These are the equations needed to predict the stepped-up or stepped-down secondary-side current and voltage values depending on the number of turns in the coils and demonstrates that the transformer may be used to raise or lower either voltage or current equivalently [31].

Necessarily, when one is changed, the other naturally adjusts to compensate so that we will always get equal values on both primary and secondary sides. Both voltage transformers and current transformers have usage in power generation and transmission, and their construction methods differ somewhat. For the sake of simplicity, we will deal mainly with current transformers.

Historically, power transformers are usually installed and built as passive devices. They don't need outside control or monitoring and operated solely as electromechanical devices. With no digital control and no connection to any network, a power transformer was beyond the reach of cyber attackers.

Nowadays, microcontrollers have led to the development of cyber-physical systems, which allow real-time management of system states and processes via internet-connected sensors and intelligent devices. While this has great benefits for operators, it has also opened the door for threat actors to remotely attack devices. Power transformers are now

commonly equipped with microcontrollers that monitor various processes and states of the device and report back to central control stations via network connections.

Power transformer failures: causes and consequences

Power transformers play a critical role in the transmission of electrical energy. Their ability to step up alternating current (AC) voltages makes it possible to transmit electricity efficiently over long distances from generation plants to end users. Therefore, understanding the various ways in which power transformers can fail is essential for developing effective monitoring and protection mechanisms.

1. External faults

A wide range of external faults can compromise the performance of a power transformer. These include overloading, system faults, overvoltage, and operation under abnormal frequency conditions [30]. Such issues originate outside the transformer and disrupt the expected electrical input. The transformer may experience excessive voltage or current or operate at a frequency that is either too high or too low. Although we do not delve deeply into external faults, it is important to recognize that these conditions can lead to, or intensify, internal failures within the transformer.

2. Internal faults

Internal faults typically occur in one of three key areas: the isolation system, the magnetic core, or the windings.

2.1 **Isolation failures**

The isolation material, usually a specially formulated oil, provides insulation under

high-voltage conditions [27]. The core and windings are submerged in this oil; all housed within an insulated enclosure. As long as the isolation material retains its polarization properties, electric fields are effectively restrained [28]. In addition to providing insulation, the oil absorbs heat generated during transformer operation. However, oil's insulation effectiveness can degrade over time due to gas formation, aging, and contamination by air, moisture, or particulates. When polarization is compromised, the core's electric field can propagate unchecked, potentially causing structural damage.

2.2 **Core faults**

The transformer core is composed of laminated sheets of thin, insulated steel that form a closed ferromagnetic loop with low reluctance. This design guides the magnetic flux, enabling efficient induction in the secondary winding. Lamination prevents the formation of inrush currents while maintaining magnetic properties. If conductive material bridges the insulated laminations or the insulation itself deteriorates, the core behaves like a solid block of steel. This allows harmful eddy currents to flow, resulting in overheating and accelerated degradation of the surrounding insulating oil.

2.3 **Winding faults**

The most common internal faults occur in the transformer windings [28]. Over time, insulation may deteriorate, leading to short circuits either between different windings (primary and secondary) or between turns of the same winding. Despite conventional illustrations, transformer windings are often coiled concentrically around the core, making such faults more likely than they might appear. Insulation failure is often driven by overheating of the dielectric oil, which weakens the winding insulation. As

contamination increases and the oil loses its polarization, arcing or direct contact may occur. Additional causes include aging materials and transient overvoltage events such as lightning strikes.

A short circuit effectively drops the voltage between the affected windings to zero, which can severely disrupt transformer operation. The resulting current may exceed the thermal limits of the conductor material, causing further issues such as conductor annealing and insulation charring. Equipment may malfunction due to voltage irregularities near the fault location. Furthermore, increased current raises the temperature of the oil, further damaging insulation and initiating a self-reinforcing cycle of failure that can ultimately lead to complete transformer breakdown.

Economic and operational impact

Transformer failures can be financially devastating. Repair or replacement is costly and time-consuming. Transformer prices vary based on several factors, such as megavolt-ampere (MVA) rating and core design [26]. A standard large transformer (≤10 MVA) may cost around $600,000, while a high-capacity or custom-designed unit can exceed $4,000,000. Given the specialized manufacturing and long lead times, it may take several months or even years to produce a replacement [23], [25]. During this period, the electrical grid must compensate for the lost capacity, leading to increased strain and potentially degraded service quality for consumers.

Power transformer protection relay

To ensure uninterrupted power supply and minimize potential damage, power engineers implement protective measures to prevent faults in transformers. A transformer

protection system plays a critical role in maintaining the reliability of power delivery by continuously monitoring key electrical variables such as current, voltage, power, and frequency to determine the operational state of the system.

These operational states may include normal functioning, degraded performance, or complete failure. Decisions about the state of the transformer are made by protective devices known as relays. A protective relay analyzes incoming data from the system and determines whether the transformer is operating within safe limits. If unsafe conditions are detected, the relay sends a trip signal to a circuit breaker, disconnecting the transformer from the grid. Although this disconnection may temporarily reduce the efficiency of power distribution, it helps with avoiding severe damage to the transformer, the substation, and surrounding infrastructure [15].

Each subsystem within a transformer is protected by its own scheme, including specific thresholds for acceptable performance. These thresholds define the conditions under which a trip should occur, and each protection scheme is designed to act independently if necessary.

Role and function of relays

A relay is an electrically operated switch that opens or closes circuits based on the conditions it monitors. A protective relay is designed specifically to detect abnormal or intolerable operating conditions within the transformer system and respond accordingly. Relays operate circuit breakers and can either allow the transformer to function normally or trigger its disconnection during fault conditions.

Relays are typically categorized into two main types:

- Measuring relays, which continuously assess and monitor variable values.
- On-off (all-or-nothing) relays, which include subtypes like time-lag, auxiliary, and tripping relays.

The tripping relay is of particular importance in transformer protection. It monitors selected system variables and compares them against predefined thresholds. When any of these values exceed safe limits, the tripping relay issues a command to isolate the transformer by activating the circuit breaker. During normal operation, these thresholds are not breached, and the system continues without interruption [18].

Evolution of relays technology

Relay technology has evolved significantly from electromechanical and solid-state designs to modern computer-based or digital relays. Our work focuses on these digital relay systems.

Digital relays offer several advantages:

- Remote monitoring and control capabilities.
- Data collection for improved operational analysis.
- Reduced need for on-site maintenance or manual data collection.

However, these benefits come with increased cybersecurity risks. Unlike older mechanical systems, digital relays are often network-connected and sometimes even accessible via the public internet. This connectivity makes them vulnerable to cyberattacks by adversaries capable of bypassing even robust security measures.

Structure of a digital relay system

A modern digital relay (Figure 3) includes the following components:

- Analog input subsystem
- Digital input and output subsystems
- Relay logic and programmable settings
- Digital signal processing filters

Input to the relay consists of both analog signals (e.g., 60 Hz voltages and currents from distribution lines) and digital signals (e.g., on/off status indicators and voltage level changes). Typically, a relay handles between 3 to 30 analog inputs and 5 to 10 digital signals. Analog signals are scaled and converted to digital form using an Analog-to-Digital Converter (ADC).

Digital output signals are sent via the relay's output interface, usually a parallel port to control external systems. The analog inputs are sampled at frequencies ranging from 240 Hz to 2000 Hz and stored in the relay's RAM as well as in its log. These records can later be accessed for diagnostics and forensic analysis.

At the core of the relay is a relaying algorithm. This algorithm processes the sampled data to determine whether the transformer is in a hazardous state. Based on this analysis, the system decides whether to maintain operation or to isolate the transformer by sending a trip signal [18].

As an example of real-world digital protection relays and automation intelligent devices (IED's), Schweitzer Engineering Laboratories (SEL) is a leading manufacturer in this field, one of the famous transformer protection relays is the SEL-787 (Figure 3), it has a set of unique features and capabilities that include a comprehensive transformer protection with differential protection for up to three-winding transformers. It also has a through-fault and overcurrent protection to guard against external faults and overloads. It

also covers advanced harmonic restraint and blocking as well as an interface for secure communications [32].

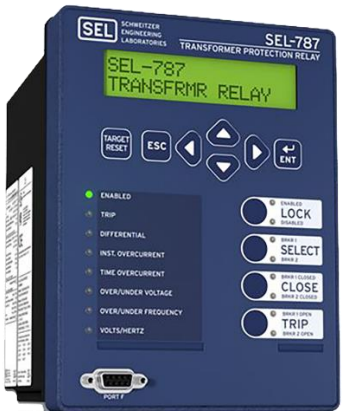


Figure 3: Transformer relay

SEL has also invented an advanced protection relay like SEL 484E and 487B (Figure 4) with a highly programmable logic engine capable of advanced automation and an interface for high speed and secure communications, It's more focused on large critical power transformers [32].

Figure 4: Advanced protection relay

Protection algorithms

Few prior research approaches addressed power transformers' protection algorithms' resilience against cyber-attacks and malware. Power engineers usually design these algorithms in the testing and modeling phases, while it is important to assess these algorithms during operation. Our focus of this research is on the differential and harmonic restraint protection algorithms since both algorithms are concerned with power transformer protection. The differential protection principle is based on comparing the output and input transformer currents, and power transformers operate with the nominal current. Current transformers (CTs) are usually selected with a corresponding turn ratio where currents in CT secondary sides are equal [33], see figure 5 below. In this case, the current will not flow when the difference between currents equals zero because secondary currents have equal amplitude and phase displacement value, so the differential current will not operate [33].

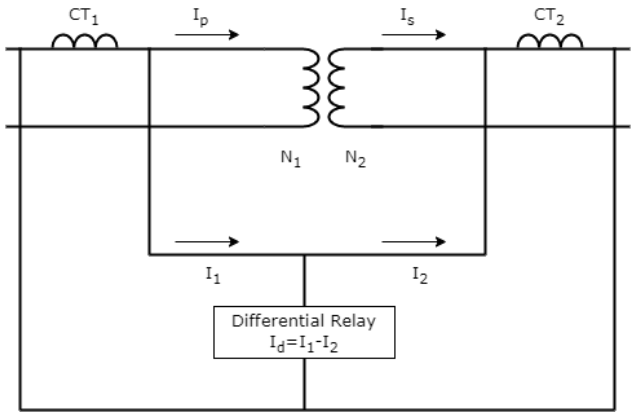


Figure 5: Differential protection for a 2-windings transformer

We will now expand more about differential and harmonic protection algorithms at one protection relay function that is attached to the power grid:

Differential protection

Differential protection is a primary method used to safeguard power transformers from internal faults. It works by comparing the current entering and leaving the transformer on each phase. Under normal conditions, these currents are equal, and the only variation arises from minor differences due to the current transformers' (CTs) magnetizing current. However, if an internal fault occurs, a significant imbalance is created between the primary and secondary currents, triggering the relay to issue a trip signal.

To facilitate accurate current comparison, CTs are installed on both the primary and secondary sides of the transformer (referred to as CT1 and CT2). These CTs scale the current proportionally to their respective turn ratios, resulting in two secondary currents:

$I1 = \frac{Ip}{N1}$, $I2 = \frac{Is}{N2}$

where Ip and Is are the primary and secondary currents, and N1, N2 are the CTs' secondary coil turn counts.

According to Kirchhoff's Current Law (KCL), these scaled currents flow in opposite directions through the relay, ideally canceling each other. In the absence of faults, the differential current $I_d = I_1 - I_2 = 0$. If I_d exceeds a predefined threshold known as the restraining current I_{rt}, the protection system interprets this as a fault and signals the circuit breakers to isolate the transformer [27].

Harmonic Restraint Protection

When a transformer is first energized, it experiences a transient condition known as magnetizing inrush. During this non-linear process, the transformer core builds up magnetic flux, causing a surge of current on the primary side, often 8-10 times the normal

peak. This inrush current does not appear on the secondary side, creating a current imbalance that may falsely trigger the differential protection.

To prevent false trips, harmonic restraint protection is used. Magnetizing inrush currents contain distinct harmonic content especially a strong second harmonic component. Using Fast Fourier Transform (FFT), the relay system analyzes the frequency spectrum of the current and checks for harmonic distortion [27].

Inrush currents typically exhibit the following harmonic profile (as a percentage of the total signal):

- Fundamental: 100%
- 2nd harmonic: ~55%
- 3rd harmonic: ~63%
- 4th harmonic: ~26.8%
- 5th harmonic: ~5.1%
- Others: smaller contributions

If the second harmonic exceeds a set threshold usually around 20% of the fundamental the relay recognizes the event as an inrush condition and suppresses tripping. Once the core is fully magnetized and the transformer reaches a steady state, harmonic levels fall, and differential protection resumes its normal fault detection [28].

Signal generation

For simulation purposes, AC signals with variable frequencies and amplitudes (e.g., 60 Hz in the U.S.) are generated to model transformer behavior. A class can be developed to synthesize sinusoidal signals and harmonics. Harmonics are defined as integer multiples

of the fundamental frequency (e.g., n=2, 3, 4,…) A composite signal is formed by summing these components.

Using a numerical library that implements the FFT algorithm, the system decomposes the signal into its harmonic components. The algorithm detects whether harmonic content, especially the 2nd harmonic, surpasses the configured thresholds, helping the relay logic distinguish between real faults and transient inrush conditions [29].

By integrating harmonic restraint with differential protection, transformers are safeguarded against both actual internal faults and benign events like inrush currents, ensuring reliable and accurate protection.

Physics-based malware detection approaches

Many proposed detection schemes in the literature aimed to address physics-based cyber-attacks targeting substation equipment in power systems, varying from behavior-based methods as in [22] to stateless and stateful traffic inspection methods as in [12] and [5]. In contrast, a few approaches studied assessing protection algorithms from an operating system perspective.

In [10], the authors proposed an approach for lowering the estimated error bound to enhance the resilience and monitor and control functionalities of energy critical systems using the Gaussian process. The authors proved that the optimal estimation algorithms can be improved using prior set inclusion constraints with prior information from the power system, and this can yield to better state estimation and optimization and probably can count toward better tight coupling of physics. This conclusion can show that better

state estimation and prior analysis of information from the power system can count toward better physics-based malware detection and prevention.

There is an accountable record of research toward limiting the impact of stealthy attacks targeting power systems and ICSs in general; few schemes focused on physics-based attack detection were noticed. For instance, the authors in [24] studied if physics-based attack detection can limit the impact of such stealthy attacks, in their study they proved that many detection schemes in the literature didn't address the problem.

They proposed a new metric to measure the impact of stealth physics-based attacks by depending on stateful tests. They showed that for a stealthy actuator attack, the integrative part of the controller corrects the system deviation and forces the attacker to have an effective impact asymptotically, concluding that launching stealth attacks in actuators is more difficult than in sensors.

Another group of researchers presented a Trusted Safety Verifier (TSV) framework [14], claiming that the safe behavior of a control system relies on a trusted computing base (TCB), where minimal code execution is allowed on commodity machines, due to the large amount of TCBs available in the market, those TCBs may contain known vulnerabilities, then minimizing the risk of executing untrusted codes is important. TSV provided test cases with translating the controller code to an intermediate language (ILIL), which allowed checking for codes containing more instructions and features than the previous controller code safety verification techniques [14].

The proposed scheme achieved reasonable efficiency by using a new hybrid symbolic execution model checking algorithms, and after evaluation, it proved that it could work as a portable device for efficient and practical verification of control programs before they are uploaded to controllers.

Machine and deep learning-based detection

Detecting physics-based cyberattacks and malware in protective relays is a complex and critical task for ensuring the security and reliability of power systems. Detecting such attacks with traditional security methods can be challenging. Therefore, with its capability of early detection, machine learning (ML) can play a significant role by leveraging its ability to analyze and uncover hidden patterns in relay data that may indicate potential malware activity.

Few proposed ML-based detection schemes in the literature aimed to address physics-based malware and cyberattacks targeting substation equipment in power systems, because the concept itself might be contained as a subtopic in the intrusion detection systems field. Usually, ML approaches can be classified into four major categories, namely, supervised Learning, unsupervised Learning, semi-supervised Learning, and reinforcement learning and according to [2], most of the intrusion detection work available in the literature is related to the first two categories.

For physics-based malware and cyberattacks detection, unsupervised learning might be preferred because it does not require labeled training data, depending on the nature of such attacks. Therefore, the dependency on attack data gets eliminated making it capable of detecting zero-day attacks [2].

Recently, behavior-based ML approaches are being used extensively to secure ICSs, depending on the operational data from the physical system, a model can be trained on the normal and abnormal behavior of the process. So, any deviation from normal behavior can be learned and observed accordingly.

In [20], the author proposed a solid approach for encountering physics-based malware by exploring the concept of active physics in the memory of an IED, the author presented OS algorithms for physics-driven page fault handling that enable deception tools to recognize and track active physics in memory, the author claimed that in realistic threat models of the electrical power grid, malware may run arbitrary code on a compromised industrial computer, moreover, malware can run in the same address space as the differential and harmonic restraint protection algorithms.

Access to physics data could be performed by code injection or by replacing the original executable files with deceptive code which would appear as running an emulated copy of the protection algorithms to deceive malware [20]. For physics recognition, the author has developed Hidden Markov Models (HMMs) and Convolutional Neural Network (CNN) layers, which are trained to recognize the physics of the power transformer and its controlled variables.

Another research study [7] employed a fully connected autoencoder for the detection and mitigation of two different cyberattack scenarios against transformer differential protective relays, the autoencoder was trained with measurements from current transformers at both sides of a transformer, then it was used to detect anomalies in

currents after activating the cyberattacks, the first attack scenario was a supply chain attack which involves the implementation of a compromised electrical unit inside the merging unit, the malicious electrical unit was able to tamper physics data received from circuit transformers, while the second scenario was an FDI where false data packets were injected onto the process bus forcing the differential relay to mis operate. In both attack scenarios, falsified physics data triggered the protective relay and caused false tripping to the power transformer without internal or physical faults. In conclusion, the simulated experiment proved the capability and high performance of the proposed ML algorithm to detect cyberattacks. The authors indicated that further research would lead to better understanding of the range of conditions where autoencoders perform well.

Challenges with machine and deep learning detection

While machine learning-based approaches have the potential to improve the resilience of digital relays, they also bring forth a unique set of challenges within this context:

1. Data quality and availability: Machine learning models require large and high-quality datasets for training and validation. Obtaining such data for cyber-physical attacks in smart grids can be challenging due to the rarity of real-world incidents. Simulated data may not accurately capture the complexities of actual attacks.
2. Adversarial attacks: Adversarial attacks, where attackers manipulate data to deceive machine learning models, are a significant concern in cybersecurity. Digital relays can be vulnerable to these attacks, which can lead to misclassification or false alarms, compromising the resilience of the system.

3. Generalization and transfer learning: Smart grids may have diverse and dynamic operating conditions. Machine learning models should be able to generalize and adapt to new and unseen attack scenarios.

CHAPTER THREE: THREAT MODEL

Traditional signature-based malware detection methods are no longer sufficient to detect new and sophisticated malware attacks. Moreover, protection relays have limited computational resources, making it challenging to implement complex malware detection methods [10]. Therefore, it's important to find robust malware detection approaches that can efficiently detect new and unknown malware attacks.

According to [11], there are multiple kinds of threats that may cause damage and disrupt power system components. These major threats are listed in table 1 below, these threats are categorized based on the level of risk behind them:

Table 1: Cyber threats against power transformers

Cyber threat	Risk level
Compromise the user-interface	Medium
Interrupting the time synchronization process	Medium
Compromise the station level communication bus	Medium to High
Gaining access to bay level devices	High
Bypassing the firewall to gain access to the substation network	Medium
Packets spoofing in channels	Medium
GOOSE messages timing attacks	Medium
ROW hammer memory attacks	Medium to high
Modifying protection relay settings	Medium
Capturing and modifying GOOSE messages	Medium
Compromising the process level communication bus	Medium to high
Placing altered values in SV messages	Medium to high
Performing FDA (False data injection) attacks	High
DDoS cyber attacks	Low to medium
Man in the middle attacks	Medium to high
Firmware-level bugs	High

From these entry points, attackers may perform different attack variations, such as message replay, command, and false data injection, and more.

The overall threat model can be summarized in figure below, then we expanded it into phases afterwards describing the steps that an attacker might follow to cause damage to a power transformer.

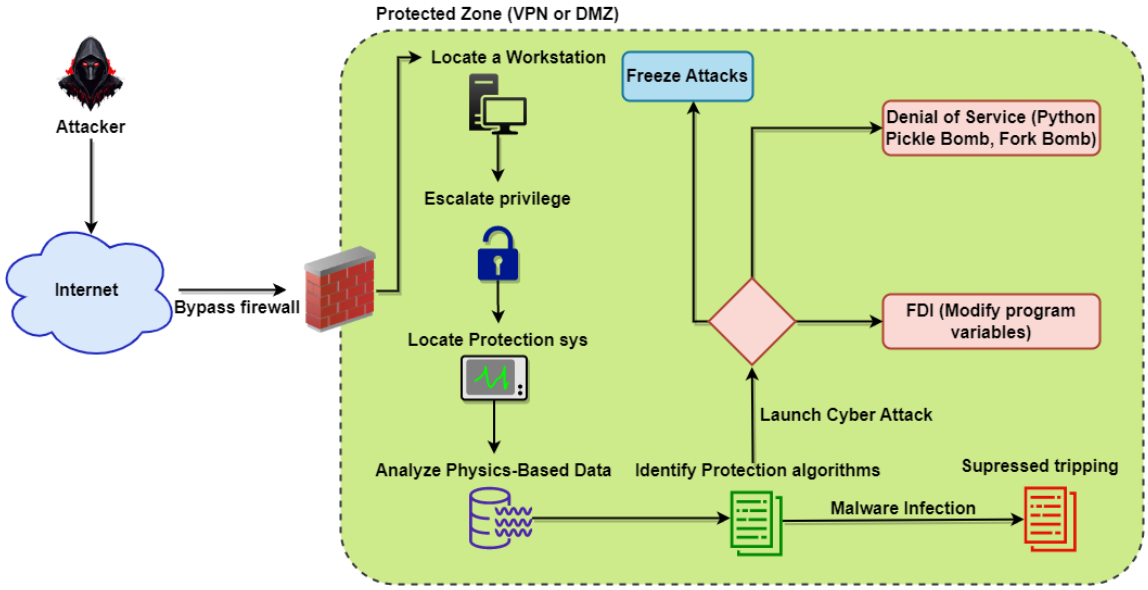


Figure 6: Threat model overview

We start our assumption for this research by stating that attackers can locate a protection relay or a command-and-control workstation where they can spy on the physical processes and determine where exactly the protection algorithms are running and infect it with malware or disrupt it in multiple ways. The attacker may bypass firewalls and intrusion detection systems for an ICS and start intercepting communication, even an attacker might be an insider or uses social engineering tricks to get and maintain access

and camouflage on the operator system with the intent of stealing information or causing damage, see figure 6 below.

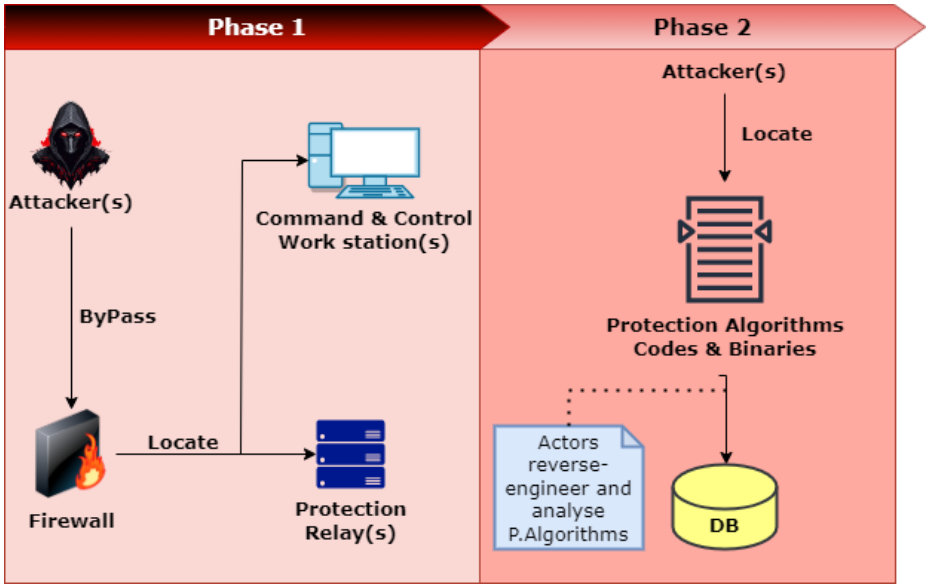


Figure 7: Threat actors' phases 1& 2

When attackers gain access to a protection relay or workstation, they may disable, bypass, or falsify protection algorithms in multiple ways causing false tripping or unexpected behavior. Therefore, maintaining integrity and normal operation of the protection algorithms is deemed very important, a protection algorithm needs to be validated and assessed following some validity checks. Additionally, some other factors can count for assuring protection algorithms correctness, estimating and tracing some performance factors during runtime can aid in detecting malicious activities or malware.

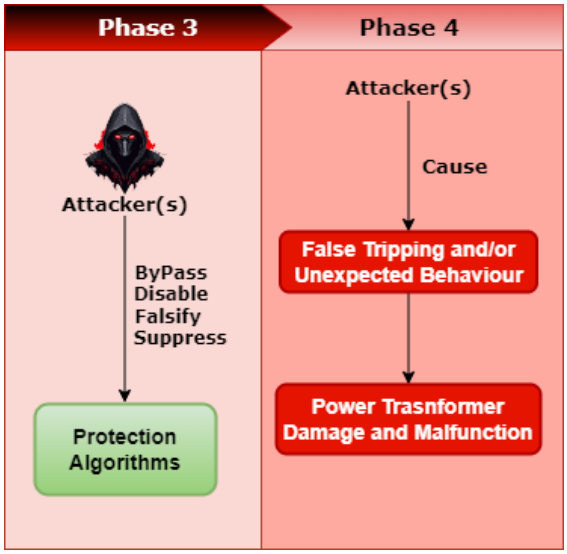


Figure 8: Threat actors' phases 3& 4

Our focus on this research is about studying the possible ways in which a protection algorithm can be deceived, and to provide a way to ensure that protection algorithms are intact and not affected by unknown malware or stealth cyberattacks. Upon that, all cyberattacks inducing physics-based damage effects on the protection relay level would be mitigated.

Performance and integrity checking have been studied and employed widely in the literature, but according to [8], few approaches were focused on assessing power transformer protection algorithms performance and resistance against cyberattacks during runtime, especially that most IDS (Intrusion detection systems) cannot handle advanced cyberattacks since they are designed to monitor a specific area.

In contrast, a robust model-based detection scheme which can be applied to the continually running protection algorithms could be an architectural advancement for detecting physics-based malware and even different kinds of malware and may exceed

that by detecting unknown intrusive codes based on assessing protection algorithm's reaction before and during runtime.

There are many proposed approaches in literature toward detecting malware through performance analysis and machine learning. For example, in [9], the authors were able to detect dynamic behavior malware through deep learning classification models, where they extracted performance counters as a timeseries from a safe and an infected virtual machine, then ten different deep learning classifiers were used after normalizing and encoding the data, the authors obtained promising results for the Multi-Channel deep convolutional neural network model (MCDCNN) with highest accuracy and least overfitting problems.

According to [10] in their evaluation, further improvement is necessary for state-of-the-art IDSs, especially to deal with Masquerading attacks. A Masquerade attack in a typical IEC 61850 substation is an injection attack but with a particular improvement where both syntax-based and anomaly-based IDSs are expected to fail in detecting it, this attack involves altering old substation control messages to mimic legitimate behavior, where attackers may benefit from some gaps in the transmission mechanism of the SqNum and StNum attributes for SV (Samples values) control messages.

Moreover, the authors concluded that the MMS communication protocol over TCP/IP is vulnerable to MITM (Man in the middle) attacks under some circumstances. Experiments for exploiting MITM attacks demonstrated the feasibility of causing physical effects via manipulating some IED parameters or attributes through injecting malicious MMS commands.

Finally, maintaining the integrity of protection algorithms codes during runtime through memory protection principles like: Copy-on-write, Zero-copy, Dirty-copy on write, are still promising techniques where the protection codes or the simulated relay can run on a separate COW-enabled virtual host like QEMU. Investigating and assessing the ability to isolate protection algorithms from a memory allocation and protection perspective is indeed another area of our research problem that we are planning to implement and evaluate accordingly.

CHAPTER FOUR: EXPERIMENT AND SIMULATION

In an earlier stage, we proposed a watchdog model for physics-based anomalies detection in digital substations; it was based on a simulated experiment in python that includes the implementation of the protection system for a 2 windings 3-phases power transformer. Differential and harmonic protection algorithms were emulated, showing continuous readings for primary and secondary currents measurements entering and leaving the power transformer under 3 different states: Normal operation, added noise, memory intensive malware.

Anomalies were considered deviations of the protection algorithms execution time around a specific threshold together with the consumed performance counters and memory along the timeline.

In this work, as we mentioned earlier, we aim to step further and focus on physics-based intrusive behavior for protection algorithms with performance metrics analysis and defensive deception or decoys and machine learning. So, we rely on the same previous model, but we extended the experiment and added and implemented new functionalities that include simulating and implementing a distress scenario, two different types of cyber-attacks and malware against the protection system and a machine learning model for detection.

To deceive attackers, we run two identical copies of the experiment, one on the local host, and the other on a virtual instance generating synthetic physics data with a conditional generative adversarial network (CGAN) model fed with various samples of normal and falsified physics data against the system so the model can learn about protection

algorithms behavior under different states of protection. Meanwhile, any access attempt targeting the decoys means that there is an incident, and all system real time behavior and intrusive events and memory intrusions or filesystem changes and even reverse shell payload injections will be identified and recorded by Fibratus tool since it can run as a service in the background [34].

CHAPTER FIVE: DETECTION APPROACH WITH MACHINE LEARNING

Machine learning (ML) and deep learning (DL) algorithms have emerged as proven methods for detecting anomalies and malware. Building on this foundation, we propose the development of an AI-driven model aimed at enhancing the detection of protection deviations in power system relays.

A thorough review of related research in the domains of malware and anomaly detection within power protection systems and relays [17][19] indicates that hybrid approaches combining machine learning and deep learning yield superior results. While ML models are highly efficient and computationally inexpensive for tasks such as data classification, DL models excel in complex pattern recognition. This synergy makes hybrid models particularly effective in security-related applications.

In addition, DL-based techniques are well-suited for detecting physics-based cyber-attacks, as these models can be pre-trained on synthetic data to reduce processing time and computational overhead. Given that our focus is primarily on classification, Support Vector Machines (SVM) are a suitable choice due to their high efficiency and accuracy.

Recently, Generative Adversarial Networks (GANs) have gained traction in the field of malware and anomaly detection, driven by the increasing complexity and sophistication of cyber threats. GANs are now being utilized across various cybersecurity applications, including enhancing Intrusion Detection Systems (IDSs) by simulating malicious traffic and behaviors [16].

For this study, we identified Conditional GANs (CGANs) as particularly relevant. These advanced models offer both offensive and defensive capabilities, making them ideal for intelligent agents and security-focused applications. In our case, CGANs are employed to emulate a defensive deception honeypot designed to attract attackers. By misleading adversaries into engaging with the honeypot, the model can learn new suppressive behavior patterns, thereby enhancing its adaptive defense mechanisms [15].

Overview about Generative Adversarial Networks (GANs)

Generative Adversarial Networks, introduced by Ian Goodfellow in 2014, are a class of deep learning models designed to generate synthetic data that closely resembles a real dataset. GANs have been successfully used to create highly realistic images, audio, and other types of data [16].

A GAN consists of two competing neural networks:

Generator (G): Produces synthetic data from random noise, with the objective of fooling the discriminator into classifying the output as real.

Discriminator (D): Distinguishes between real data (from the training dataset) and synthetic data (from the generator), attempting to correctly identify whether input samples are genuine or fabricated.

The training process is adversarial: the generator iteratively improves its ability to produce realistic outputs, while the discriminator becomes better at detecting fakes. Over time, the generator's outputs become increasingly indistinguishable from real data. (See Fig. 9 below for illustration.)

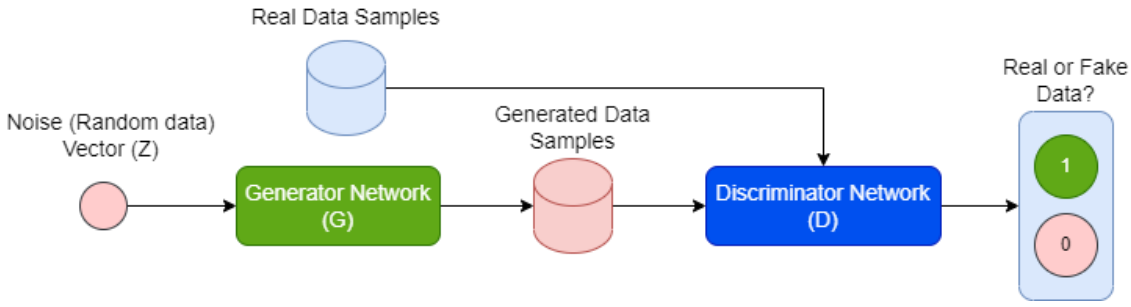


Figure 9: GAN architecture

To construct the dataset, we began by profiling the CPU execution time, memory usage, and memory allocation of the protection algorithms during normal operation. This was achieved using Python's standard libraries cprofile, psutil, and tracemalloc to capture performance metrics over time. These baseline measurements serve as a reference point, enabling the model to later assess deviations and identify abnormal behavior based on extracted features.

Next, we simulated a distress scenario that triggered abnormal behavior in the protection algorithms. Data samples were collected during this phase to represent atypical but non-malicious system states. Following this, we executed a malware scenario along with other proposed cyber-attack simulations, collecting additional samples to capture the system's response under malicious conditions.

Data samples acquisition

To construct the dataset, we began by profiling the CPU execution time, memory usage, and memory allocation of the protection algorithms during normal operation. This was achieved using Python's standard libraries cprofile, psutil, and tracemalloc to capture performance metrics over time. These baseline measurements serve as a reference point,

enabling the model to later assess deviations and identify abnormal behavior based on extracted features.

Next, we simulated a distress scenario that triggered abnormal behavior in the protection algorithms. Data samples were collected during this phase to represent atypical but non-malicious system states. Following this, we executed a malware scenario along with other proposed cyber-attack simulations, collecting additional samples to capture the system's response under malicious conditions.

Table 2: Concerned features

Name	Description
Differential algorithm	
Execution time_diff	Processor execution time required in milliseconds for the differential algorithm
Consumed memory_diff	Memory consumption by the differential algorithms in KB
Allocated memory_diff	Allocated memory by the differential algorithms in KB
Harmonic algorithm	
Execution time_har	Processor execution time required in milliseconds for the harmonic algorithm
Consumed memory_har	Memory consumption by the harmonic algorithms in KB
Allocated memory_har	Allocated memory by the harmonic algorithms in KB

Features extraction and data labeling

In this experiment, we analyzed the specified features related to the execution resource metrics of both protection algorithms. These features were collected in real time as structured, tabular raw data. Since the data was already in a clean and readable format, no additional preprocessing such as filtering or cleaning was necessary, making it suitable for direct use in machine learning models.

The dataset was organized into two-dimensional vector groups: input features (x) and corresponding labels (y), then split into training and testing subsets. Prior to model training, the data was scaled to ensure compatibility and improve learning efficiency.

For classification, five distinct labels were defined based on the operational state observed during the experiments: Normal, Distress, Malware, Cyber-Attack 1, and Cyber-Attack 2. These labels represent different behavioral states of the system and act as the target outputs for the classification model.

Support vector machine (SVM) classifier

To classify system states: normal operation, distress, malware activity, and cyber-attacks we employed a Support Vector Machine (SVM) classifier, utilizing system performance and memory-related features as inputs. SVM was selected due to its robust ability to maximize the decision margin between classes, thereby enhancing generalization performance and reducing the risk of misclassification.

To handle the inherent non-linear patterns in the data, we utilized a Radial Basis Function (RBF) kernel, which effectively maps the input space into a higher-dimensional feature space. Model hyperparameters, such as the penalty parameter (C) and the kernel coefficient (γ), were fine-tuned using grid search combined with 5-fold cross-validation. This strategy ensures optimal model configuration while minimizing the risk of overfitting by validating performance across multiple data subsets.

After determining the optimal hyperparameter set, the final SVM model was trained on the full training dataset and subsequently evaluated on a separate hold-out test set to assess generalization.

Given a training dataset (x_i, y_i), where $x_i \in R^n$ and $y_i \in \{-1, +1\}$, the SVM solves the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w} \cdot x_i + b) \geq 1, \quad \forall i \tag{1}$$

Where w is the weight vector and b is the bias.

The RBF (Radial Basis Function) kernel is a similarity function used by SVMs to handle non-linear decision boundaries. The RBF kernel is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{2}$$

Where:

x_i, x_j are two data points (feature vectors), $\| x_i - x_j \|^2$ is the squared Euclidean distance

ϒ (gamma) is a parameter that defines how far the influence of a single training example reaches

Once the scaled dataset was provided to the model, the SVM classifier was trained to identify deviations from normal behavior by learning the patterns associated with each labeled category. The output labels were encoded numerically as follows: 0 for Normal, 1 for Distress, 2 for Malware, and 3 and 4 for Cyber-Attack 1 and Cyber-Attack 2, respectively.

To visualize the dataset and classifier results, Principal Component Analysis (PCA) was applied as a dimensionality reduction technique. Since the feature space consists of six dimensions, PCA was used to project the data into two principal components PCA1 and PCA2 while preserving as much of the original variance as possible. This 2D projection enabled visualization of the feature clusters and allowed us to plot both the true labels and the predicted labels for comparison.

As shown in Figure 10, the predicted samples are clearly separated in the PCA space, aligning with the assigned class labels. In Figure 10, the classification results demonstrate that all samples were correctly classified by the model, with no misclassifications. Furthermore, as depicted in Figure 11, no significant deviations or inconsistencies were observed in the classifier's behavior.

Figure 10: PCA components 1 and 2

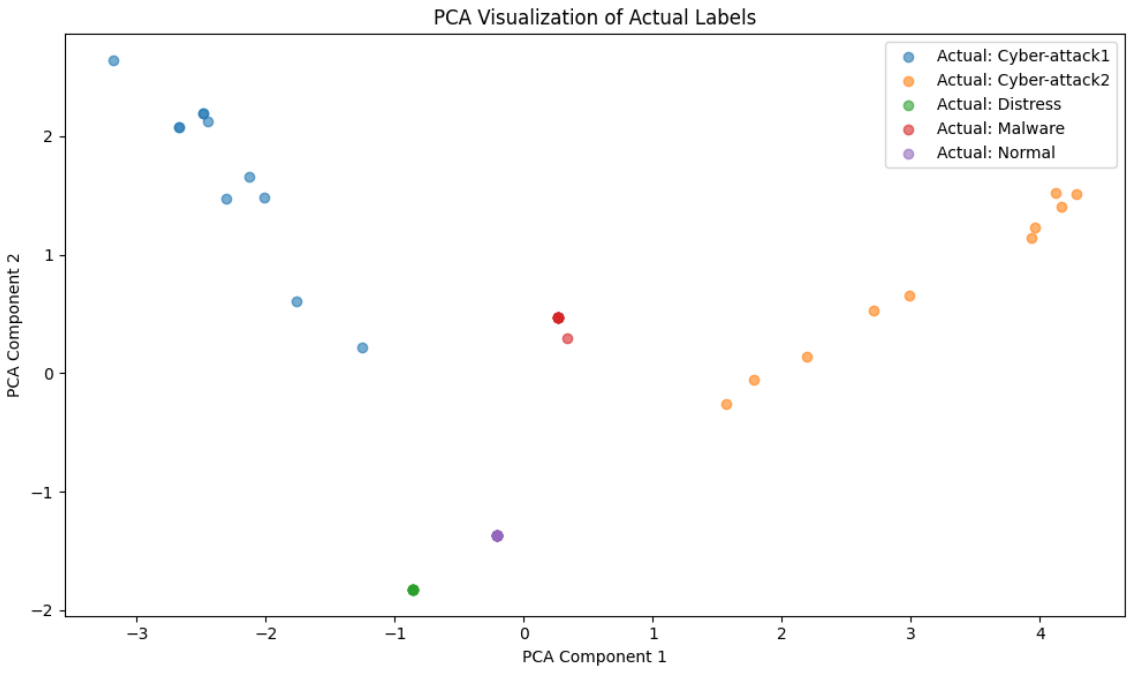
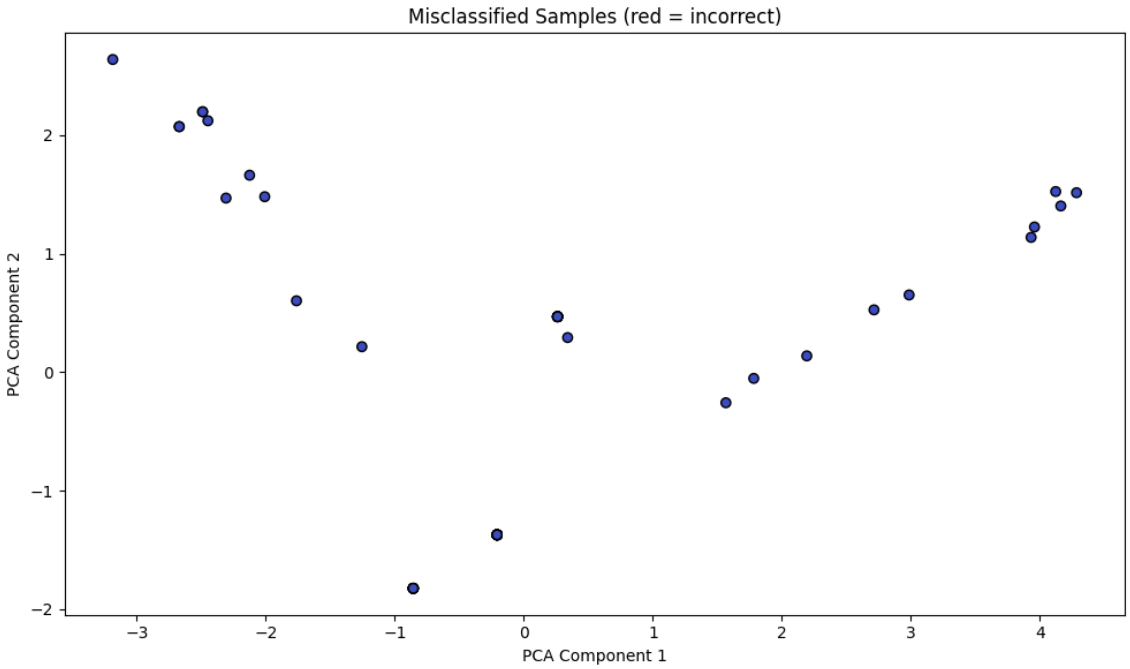
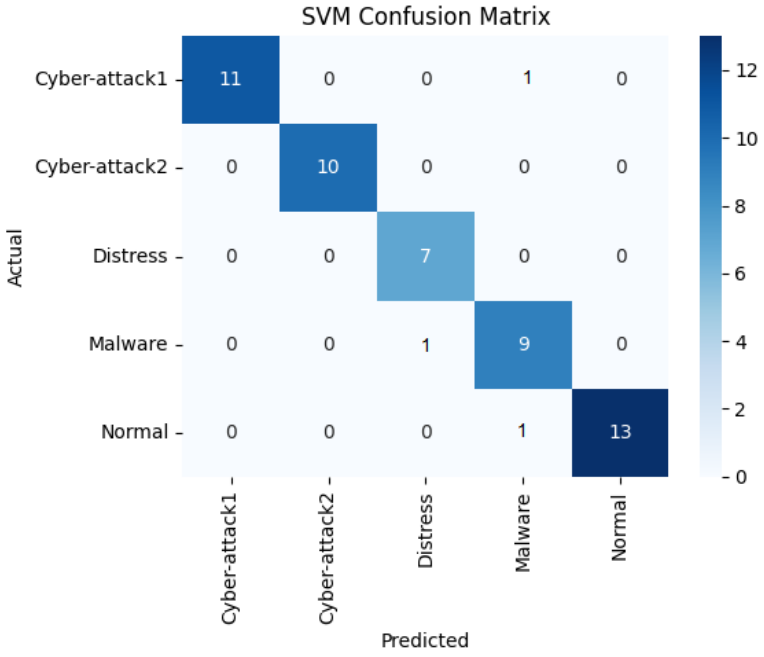


Figure 11: Samples misclassification with PCA components



After the classification phase, we noticed that all predicted labels previously mentioned are included in the confusion matrix below (Figure 12), the predicted labels according to the provided features in the dataset are almost identical to the actual labels.

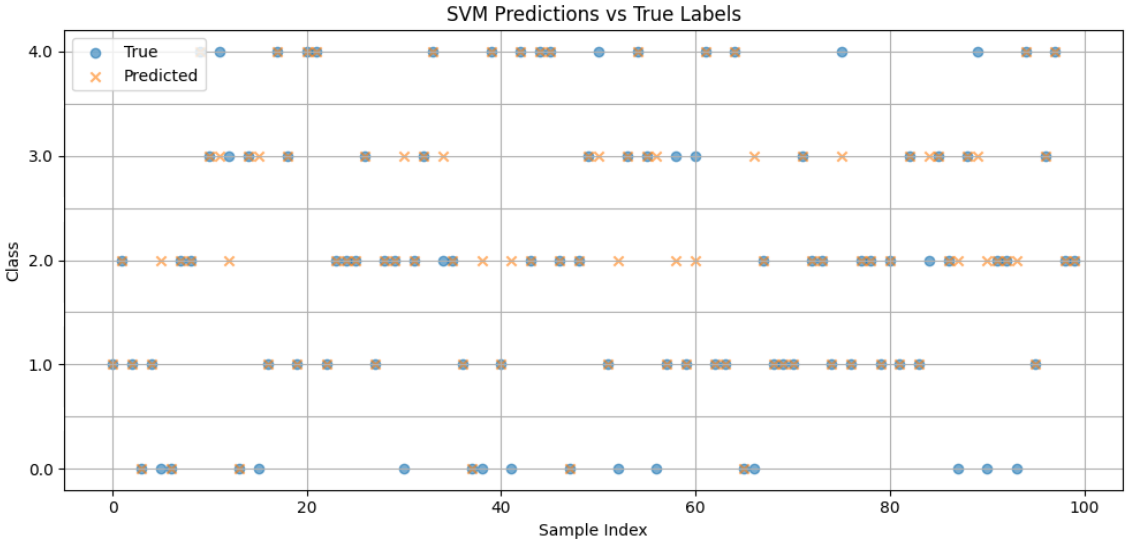
Figure 12: SVM confusion matrix



Conditional generative adversarial network (CGAN)

As previously discussed, the CGAN model generates synthetic data samples conditioned on class labels, producing six-feature vectors. The discriminator then tries to distinguish these synthetic samples from real data. A total of 300 synthetic records were generated using the CGAN to train an SVM classifier. We used a scikit-learn SVM with an RBF kernel, training it on a mix of real and synthetic data. As shown in Fig. 13, the SVM successfully identified potential protection deviations in a test set of 100 samples, specifically within classes 2 and 3 representing malware and cyber-attack type 1. These deviations suggest that both scenarios exhibit abnormal consumption of processing time and memory resources, exceeding the thresholds of normal system behavior.

Figure 13: SVM-GAN predictions



CHAPTER SIX: EVALUATION AND TESTING

This section presents the results of our proposed detection model based on the SVM-CGAN framework. We evaluated the model's performance using several key metrics to assess its accuracy and effectiveness. A summary of the results is provided in Table 4.

Table 3: Evaluation results

Metric	Accuracy	Precision	Recall	F1 Score
SVM	0.97	0.93	0.91	0.93
SVM - CGAN	0.91	0.89	0.93	0.91

Our detection model accurately identified 50 anomalous samples classified by their true labels and proximity to the decision boundary within the emulated protection algorithms. This strong performance is reflected in the model's high accuracy, precision, and recall scores, demonstrating its effectiveness in distinguishing normal from intrusive behavior and its resilience against hacking attempts.

As shown in Table 4, the SVM classifier achieved an accuracy of 0.97, clearly separating normal and suppressive behaviors. A precision score of 0.93 indicates a low rate of false positives, while a recall of 0.91 confirms the model's ability to detect true anomalies. The F1 score of 0.93, balancing both precision and recall, further underscores the model's reliability.

CHAPTER SEVEN: CONCLUSION

In this study, we developed a hybrid machine learning model designed to effectively detect protection deviations in power transformer protection algorithms. The inspiration for this work stemmed from the growing concerns surrounding industrial malware such as Stuxnet and its implications for the security of industrial control systems.

Our objectives were twofold: first, to advance our previous research on anomaly detection by introducing a machine learning model capable of addressing real-world adversarial threats; and second, to improve detection accuracy and system resilience through the integration of machine learning with generative AI techniques. This approach offers a novel perspective on securing critical infrastructure by combining adversarial machine learning with traditional learning models.

We conducted a thorough and rigorous evaluation to assess the robustness and performance of our model under various adversarial scenarios. The results underscore the urgent need for proactive defense mechanisms against sophisticated attacks. Our findings demonstrate the potential of combining deep learning and generative AI in this domain and aim to inspire the cybersecurity community to explore and adopt more advanced techniques to safeguard critical systems essential to societal infrastructure.

REFERENCES

[1] El Mrabet, Z., Kaabouch, N., El Ghazi, H., (2018). Cyber-security in smart grid: Survey and challenges. Computers & Electrical Engineering, 67, 469-482. ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2018.01.015

[2] Mendel, J. (2017). Smart grid cyber security challenges: Overview and classification. e-mentor, 68(1), 55-66, http://dx.doi.org/10.15219/em68.1282

[3] Gunduz, M. Z., & Das, R. (2020). Cyber-security on smart grid: Threats and potential solutions. Computer networks, 169, 107094

[4] Yang, Y., Littler, T., Sezer, S., McLaughlin, K., & Wang, H. F. (2011, December). Impact of cyber-security issues on smart grid. In 2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (pp. 1-7). IEEE

[5] Khoei, T. T., Slimane, H. O., & Kaabouch, N. (2022). A comprehensive survey on the cyber-security of smart grids: Cyber-attacks, detection, countermeasure techniques, and future directions. arXiv preprint arXiv:2207.07738

[6] Baumeister, T. (2010). Literature review on smart grid cyber security. Collaborative Software Development Laboratory at the University of Hawaii, 650

[7] Otuoze, A. O., Mustafa, M. W., & Larik, R. M. (2018). Smart grids security challenges: Classification by sources of threats. Journal of Electrical Systems and Information Technology, 5(3), 468-483

[8] Goel, S., Hong, Y., Papakonstantinou, V., Kloza, D., Goel, S., & Hong, Y. (2015). Security challenges in smart grid implementation. Smart grid security, 1-39

[9] Berghout, T., Benbouzid, M., & Muyeen, S. M. (2022). Machine learning for cybersecurity in smart grids: A comprehensive review-based study on methods, solutions, and prospects. International Journal of Critical Infrastructure Protection, 38, 100547

[10] Lanotte, R., Merro, M., Munteanu, A., & Viganò, L. (2020). A formal approach to physics-based attacks in cyber-physical systems. ACM Transactions on Privacy and Security (TOPS), 23(1), 1-41

[11] Urbina, D. I., Giraldo, J., Cardenas, A. A., Valente, J., Faisal, M., &Sandberg, H. (2016). Survey and new directions for physics-based attack detection in control systems (pp. 16-010). Gaithersburg: US Department of Commerce, National Institute of Standards and Technology

[12] Chen, C., Zhao, J., Zhang, K., Liu, Y., & Chen, Y. (2023). Hybrid Physics-Based and Data-Driven Mitigation Strategy for Automatic Generation Control Under Cyber Attack. In Power Systems Cybersecurity: Methods, Concepts, and Best Practices (pp. 135-160). Cham: Springer International Publishing

[13] Chen, C., Zhao, J., Zhang, K., Liu, Y., & Chen, Y. (2023). Hybrid Physics-Based and Data-Driven Mitigation Strategy for Automatic Generation Control Under Cyber Attack. In Power Systems Cybersecurity: Methods, Concepts, and Best Practices (pp. 135-160). Cham: Springer International Publishing

[14] Lanotte, R., Merro, M., Munteanu, A., & Viganò, L. (2019). A formal approach to physics-based attacks in cyber-physical systems (extended version). arXiv preprintarXiv:1902.04572

[15] Gu, Q., Formby, D., Ji, S., S. formaggio, B., Bourgeois, A., & Beyah, R. (2021). This hacker knows physics: Device physics aware mimicry attacks in cyber-physical systems.IEEE Transactions on Dependable and Secure Computing, 19(5), 3218-3230

[16] Liu, C., Alrowaili, Y., Saxena, N., & Konstantinou, C. (2021). Cyber risks to critical smart grid assets of industrial control systems. Energies, 14(17), 5501

[17] Macaulay, T., & Singer, B. L. (2011). Cybersecurity for industrial control systems: SCADA, DCS, PLC, HMI, and SIS. CRC Press

[18] Zhang, F. (2019). Cybersecurity solutions for industrial control systems and key equipment

[19] Enescu, F. M., Bizon, N., & Moraru, C. M. (2019). Issues in Securing Critical Infrastructure Networks for Smart Grid Based on SCADA, Other Industrial Control and Communication Systems. Power Systems Resilience: Modeling, Analysis and Practice,289-324

[20] Lankatilake, K. K. T. (2019). A taxonomy-based analysis of attacks on industrial control systems

[21] Teixeira, A., Kupzog, F., Sandberg, H., & Johansson, K. H. (2015). Cyber-secure and resilient architectures for industrial control systems. In Smart Grid Security (pp. 149-183). Syngress

[22] Julian L. Rrushi. 2022. Physics-Driven Page Fault Handling for Customized Deception against CPS Malware. ACM Trans. Embed. Comput. Syst. 21, 3, Article 23 (May 2022), 36 pages. https://doi.org/10.1145/3502742

[23] Ruben, Cody and Dhulipala, Surya and Nagaraj, Keerthiraj and Zou, Sheng and Starke, Allen and Bretas, Arturo and Zare, Alina and McNair, Janise “Hybrid data-driven physics model-based framework for enhanced cyber-physical smart grid security”, IET Smart Grid 2020

[24] Olijnyk, B. Bond and J. Rrushi, "Design and Emulation of Physics-Centric Cyberattacks on an Electrical Power Transformer," in IEEE Access, vol. 10, pp. 15227-15246, 2022, doi: 10.1109/ACCESS.2022.3148046

[25] Creating Effective Industrial-Control-System Honeypots Neil C. Rowe, Thuy D. Nguyen, Marian M. Kendrick, Zaki A. Rucker, Dahae Hyun, and Justin C. Brown, Proceedings of the 53rd Hawaii International Conference on System Sciences 2020

[26] Nankya, M.; Chataut, R.; Akl, R. Securing Industrial Control Systems: Components, Cyber Threats, and Machine Learning-Driven Defense Strategies. Sensors 2023, 23, 8840. https://doi.org/10.3390/s23218840

[27] Rehaan Irani. “What is Alternating Current?” Circuit Basics, https://www.circuitbasics.com/what-is-alternating-current/ (accessed on May 10, 2021)

[28] J. B. Calvert. “Inside Transformers.” University of Denver. https://web.archive.org/web/20070509111407/http://www.du.edu/~jcalvert/tech/transfor.htm (accessed April 23, 2021)

[29] William Eccles. 2011. Pragmatic Electrical Engineering: Fundamentals. Synthesis Lectures on Digital Circuits and Systems 6, 1 (2011), 1–199. J. ACM, Vol. 37, No. 4,Article 111. Publication date: August 2021

[30] Geeta Yadav, Kolin Paul, Architecture and security of SCADA systems: A review, International Journal of Critical Infrastructure Protection, Volume 34, 2021, 100433, ISSN 1874-5482, https://doi.org/10.1016/j.ijcip.2021.100433

[31] H.Tarazi, S.Sutton, J.Olinjyk, B.Bond, J.Rrushi, A watchdog model for physics-based anomaly detection in digital substations, International Journal of Critical Infrastructure Protection, Volume 44 2024, 100660, ISSN 1874-5482, https://doi.org/10.1016/j.ijcip.2024.100660

[32] Schweitzer Engineering Laboratories, A Three-Phase Transformer Protection, Automation, and Control System, SEL-487E-3, -4 Data Sheet, Schweitzer Engineering, 2021, URL: https://cdn.selinc.com/assets/Literature/Product%20Literature/Data%20Sheets/487E-3-4_DS_20210708.pdf?v=20210723-164841

[33] M. Hawary. 2008. Introduction to electrical power systems. IEEE Press Series on Power Engineering, Vol. 50. Wiley. https://books.google.com/books?id=eLWh9w8zZpIC

[34] Fibratus tool, Realtime behavior detection, memory scanning, and forensics capabilities, https://www.fibratus.io/#/overview/what-is-fibratus by Nedim Šabić