
Article:
Cali, Umit, Catak, Ferhat Ozgur and Halden, Ugur (2024) Trustworthy cyber-physical power systems using AI:dueling algorithms for PMU anomaly detection and cybersecurity. Artificial Intelligence Review. 183. ISSN: 0269-2821

https://doi.org/10.1007/s10462-024-10827-x

Artificial Intelligence Review (2024) 57:183
https://doi.org/10.1007/s10462-024-10827-x

Trustworthy cyber-physical power systems using AI: dueling algorithms for PMU anomaly detection and cybersecurity

Umit Cali[1,3] · **Ferhat Ozgur Catak**[2] · **Ugur Halden**[3]

Accepted: 6 June 2024 / Published online: 21 June 2024

Abstract
Energy systems require radical changes due to the conflicting needs of combating climate change and meeting rising energy demands. These revolutionary decentralization, decarbonization, and digitalization techniques have ushered in a new global energy paradigm. Waves of disruption have been felt across the electricity industry as the digitalization journey in this sector has converged with advances in artificial intelligence (AI). However, there are risks involved. As AI becomes more established, new security threats have emerged. Among the most important is the cyber-physical protection of critical infrastructure, such as the power grid. This article focuses on dueling AI algorithms designed to investigate the trustworthiness of power systems' cyber-physical security under various scenarios using the phasor measurement units (PMU) use case. Particularly in PMU operations, the focus is on areas that manage sensitive data vital to power system operators' activities. The initial stage deals with anomaly detection applied to energy systems and PMUs, while the subsequent stage examines adversarial attacks targeting AI models. At this stage, evaluations of the Madry attack, basic iterative method (BIM), momentum iterative method (MIM), and projected gradient descend (PGD) are carried out, which are all powerful adversarial techniques that may compromise anomaly detection methods. The final stage addresses mitigation methods for AI-based cyberattacks. All these three stages represent various uses of AI and constitute the dueling AI algorithm convention that is conceptualised and demonstrated in this work. According to the findings of this study, it is essential to investigate the trade-off between the accuracy of AI-based anomaly detection models and their digital immutability against potential cyberphysical attacks in terms of trustworthiness for the critical infrastructure under consideration.

1 Introduction

Modern power systems are evolving at a pace like never before. This significant transformation is primarily driven by five key elements, commonly referred to as the '5 Ds' of energy: Deregulation, Decentralization, Decarbonization, Digitalization, and Democratization. Among these, Decarbonization and Digitalization are the most influential trends shaping the future of our power grids (Cali et al. 2021). Decarbonization is the process of decreasing the amount of greenhouse gas emissions generated by the burning of fossil

Extended author information available on the last page of the article

fuels. Due to climate change and rising energy demand, the energy industry has faced unprecedented challenges in recent years. Energy production is a major source of greenhouse gas emissions, hence it is essential for reducing the consequences of climate change. In order to attain net-zero emissions by 2050, the European Union (EU) has set aggressive goals for lowering greenhouse gas emissions by switching to renewable energy sources (RES) (https://www.ipcc.ch/sr15/chapter/spm/). Additionally, it entails promoting energy efficiency and advancing technology for the collection and storage of carbon dioxide. The primary objective of decarbonization is to establish a sustainable and ecologically sound energy system. Nevertheless, it is crucial to acknowledge that decarbonization is a multifaceted procedure that requires substantial transformations in energy generation, infrastructure, and consumption patterns. However, similar yet distinctly different changes have been occurring within the energy industry, starting from the global Organization of Petroleum Exporting Countries (OPEC) crisis 40 years ago. Thus, in order to better protect the energy industry against sudden shocks, an evolution within the energy industry was deemed necessary and has been realized via various policy changes initiated and led by governments across the globe. The deregulation of the energy industry led to alternative utility models compared to the traditionally integrated utility model (Karney 2019), which started the transformation from large and centralized energy systems towards smaller and distributed systems that collaborate with each other (Bauknecht et al. 2020). This policy change led to the decentralization of the energy systems, which resulted in higher RES deployment and utilization, in combination with various governmental policy schemes such as Feed-in-Tariffs, which transformed the energy system into decarbonization. Furthermore digitalization is shaping rapidly decarbonized and decentralized power system. High utilization of distributed RES, such as residential-scale PV systems, resulted in new challenges while managing the power system, such as regional imbalances, supply and demand issues within certain regions, etc. Thus, as a solution to these challenges, smart grid systems were proposed that utilize advanced Information and Communication Technologies (ICTs), the Internet of Things (IoT), and various Artificial Intelligence (AI) techniques. These changes working in tandem, led to the digitalization of the energy sector while also enabling new paradigm shifts such as Peer-to-Peer (P2P) energy trading and hence, resulting in the democratization of the energy sector. With advancements in technology and the integration of intelligent devices, the energy sector has witnessed a paradigm shift towards more efficient, reliable, and sustainable power systems. The digitalization of energy systems has emerged as a transformative force, revolutionizing how power is generated, transmitted, and consumed. As part of this digital transformation, the application of artificial intelligence (AI) in power systems has gained significant momentum, enabling enhanced decision-making, automation, and optimization within the power sector. Digitalization unveils both advantages and disadvantages in the power markets and systems. On one hand, it offers new opportunities to enhance effectiveness and efficiency of the power markets and systems. On the other hand, it also introduces additional risks to cyberphysical security, as it creates more vulnerable surfaces that potential attacks or interventions from outside actors can target. For instance, anomaly detection plays a crucial role in maintaining the integrity and stability of energy systems by identifying abnormal behaviors or events that deviate from expected patterns. These anomalies could range from equipment malfunctions and cyberattacks to natural disasters and human errors. Timely detection of such anomalies is vital for preventing disruptions, minimizing downtime, and ensuring the security and reliability of power systems. AI techniques, such as machine learning and deep learning, have been leveraged to address various challenges in power systems, including load forecasting, fault diagnosis, demand response, and anomaly detection.

Traditional anomaly detection techniques, such as statistical methods, rule-based approaches, and expert systems, have been widely used in energy systems. These methods often rely on pre-defined thresholds or rules to flag abnormal events. However, they may struggle to capture complex, dynamic anomalies that evolve or exhibit subtle variations. With the advent of AI and machine learning, more sophisticated anomaly detection algorithms have been developed and applied in energy systems. Machine learning techniques, including supervised, unsupervised, and semi-supervised learning, have shown promise in detecting anomalies by learning patterns and anomalies directly from data. Deep learning approaches, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have also gained traction in anomaly detection tasks, leveraging their ability to model complex temporal and spatial dependencies in energy system data. However, as the integration of AI algorithms becomes more prevalent in energy systems, the issue of cybersecurity has come to the forefront. The vulnerabilities associated with the digitalization of energy systems and the increasing reliance on AI present new challenges and risks that must be addressed effectively. The convergence of cybersecurity and AI introduces unique considerations and potential adversarial threats that can compromise the robustness and effectiveness of anomaly detection algorithms in energy systems.

Cybersecurity issues in energy systems arise from the interconnectedness of devices, the exposure to external networks, and the potential for malicious attacks targeting critical infrastructure. The power sector, being a prime target for cyber attacks due to its importance and interdependencies with other sectors, must proactively address these challenges to ensure the integrity and resilience of energy systems. The use of AI in power systems introduces additional concerns, as AI algorithms can be vulnerable to various types of attacks, including adversarial ones. Adversarial attacks against anomaly detection algorithms in energy systems aim to deceive or manipulate AI models by exploiting their vulnerabilities. The objective is to introduce subtle changes to the input data, leading the anomaly detection algorithms to misclassify overlook potential anomalies, thereby undermining their reliability and effectiveness. These attacks can take various forms, such as data poisoning, evasion attacks, or adversarial examples. This article explores the implementation of dueling AI/ML algorithms designed to evaluate the trustworthiness of power systems' cyber-physical security under various scenarios using the Phasor measurement units (PMU) as use case. Particularly in PMU operations, the focus is on areas that manage sensitive data vital to power system operators' operations where we will delve into the vulnerabilities in anomaly detection algorithms when subjected to adversarial attacks. By understanding these issues, the transmission system operators (TSOs), as well as distribution system operators (DSOs), can develop robust countermeasures to enhance the resilience of anomaly detection systems in energy systems, ensuring the security and reliability of power infrastructure in the face of emerging cybersecurity threats. Therefore, the contributions of this article can be noted as: 1. Introduction of cyber-physical and social systems within the context of energy and cyberphysical security, 2. Methodological literature review, 3. Demonstration of AI-based anomaly detection algorithm for PMU use case as basis, 4. Evaluation of the implemented algorithms according to their robustness towards adversarial AI-based attacks as dueling algorithms for cyber attack and mitigation scenarios, and 5. Evaluation of the degree of trustworthiness for the investigated power systems related use case and scenarios.

The remainder of this article is organized as follows: Sect. 2 provides an overview of the digitalization of energy systems and the role of AI in power systems and cybephysical security. Section 3 discusses the cyberphysical security issues that arise in energy systems and PMUs. Section 4 focuses on the cybersecurity challenges related to AI and its impact

on anomaly detection algorithms. Finally, Sect. 5 concludes the article by summarizing the essential findings and outlining potential directions for future research in securing energy systems against adversarial attacks.

2 Related work

This article focuses on dueling AI/ML algorithms crafted to conceptualize and demonstrate the trustworthiness of cyber-physical system security under diverse scenarios using the PMU use case. The initial section deals with anomaly detection applied to energy systems and PMU-related tasks, while the subsequent section examines adversarial attacks targeting AI/ML models. The final section addresses mitigation methods for AI/ML-based cyberattacks. Notably, the authors highlight a gap in existing research, emphasizing that the impact of adversarial attacks on anomaly detection with PMUs still needs to be explored. Consequently, the authors deemed conducting a dedicated literature review on this subject imperative.

The modern power systems emerged from the more fundamental twentieth-century structure of one-way flow from centralized power generators to customers. The current grid complexity, which incorporates renewable energy sources, has enhanced, and poses challenges for conventional forecast and control techniques. ICT is a crucial aspect of the smart grid, enhancing power system reliability through intelligent infrastructure and various technologies. However, its vulnerability to failures and cybersecurity issues can compromise this reliability (Jimada-Ojuolape and Teh 2020). AI/ML has already demonstrated its effectiveness in other technical domains. These technologies may be used to improve energy forecasting, enable predictive maintenance, implement AI-driven control, and boost cybersecurity in power systems (Cali et al. 2021). Moreover, some studies have explored additional dimensions of power systems reliability and cybersecurity, including dynamic thermal line rating within the framework of cyberphysical power systems (Lawal and Teh 2023; Lawal et al. 2024). The use of AI/ML for anomaly detection in power systems has emerged as a rapidly evolving field of study. Researchers are using AI to identify anomalies in several industries associated with energy. These include many tasks, such as detecting anomalies in photovoltaic systems, batteries, PMUs, monitoring anomalies in energy use, studying power electronics, implementing advanced electric metering, and conducting predictive maintenance on power system assets, among other use cases (Amini et al. 2022; Ogu et al. 2021; De Benedetti et al. 2018; Baker et al. 2023; Himeur et al. 2021; Zhang et al. 2022; Gaggero et al. 2020, 2022). Furthermore, Ahmed et al. (2016) extensively examines the prominent anomaly detection techniques, including classification, statistical analysis, information theory, and clustering, that are used to discover network intrusions. It also explores the challenges encountered when working with datasets specifically created for this purpose.

Among the wide variety of AI-based anomaly detection for PMU-related use cases, this is one of the most promising and impactful domains since such anomalies can dramatically impact the entire power system. The high-frequency nature of PMU measurements made it possible to achieve real-time monitoring and management of electrical systems. PMUs send data to distributed substations across the system for system-wide monitoring and control, which is crucial for various power system applications such as state estimations and various anomaly detection (Veerakumar et al. 2023). However, applying stand-alone methods, such as the ones with fixed parameters for anomaly detection, takes great effort in the

tuning phase and does not yield the best results. Thus, ML applications on power system anomaly detection by utilizing time series PMU data have seen rapid research interest over the years (Halden et al. 2022). One such research was performed by Zhou et al. (2018) where the authors developed and compared various ML techniques for anomaly detection with PMU data. In total, four ML techniques (ensemble, regression, dbscan and chebyshev) were assessed for their anomaly detection performance in cooperation with various other factors such as scalability and computational power requirements. According to the results, the ensemble-based ML technique outperforms other performed techniques by a recall score of 0.92, whereas the lowest accuracy was observed during DBSCAN utilization, with a recall score of 0.86. Another research that tried to identify physical fault events such as voltage sag, sustained interruption, and under or over-voltage events was performed by Jamei et al. (2017). The authors utilized distributed Micro-PMU data in combination with a specially developed algorithm called the two-sided Cumulative Sum algorithm. The algorithm was utilized in a simulation environment with the IEEE-34 bus bar test case. According to the results, the authors identified the fault events with a total accuracy score of 96%. Nevertheless, as modern-day power grids are cyber-physical and social systems, anomalies can happen at any of those levels.

Cyber-physical security of power systems as critical infrastructure shall be investigated by considering different contents other than anomaly detection in terms of fault detection. One such example was researched by Ford et al. (2014), which proposed an ANN-based intrusion detection system in order to predict the consumption behavior of grid customers better. The authors utilized assessed customers' energy consumption behavior profiles in addition to hot encoded time data such as day of the week and weekend vs. weekdays in order to model the customers' typical consumption behavior where the statistical analysis between the real usage vs. estimated usage can be used to identify deviations from stable power grid operation. The study by Valdes et al. (2016) looked at energy measurement samples and used self-organizing maps and adaptive resonance theory to find new information and patterns that were the same. In Ashrafuzzaman et al. (2018), stealthy false data insertion in a state estimation was detected using both supervised and unsupervised machine learning techniques where dimensionality reduction is accomplished using PCA, and a distributed SVM is utilized to distinguish between a stealth assault and a regular attack. Meanwhile, Hink et al. (2014), Badrinath Krishna et al. (2016), Badrinath Krishna et al. (2015), O'Toole et al. (2019) have extensively worked with anomaly detection concerning electrical meter frauds. In Hink et al. (2014), researchers investigated an ARIMA forecasting tool and came up with a way to find strange patterns in data about how much electricity is used. Whereas, in Badrinath Krishna et al. (2015) the authors proposed a framework based on KullbackLeibler Divergence (KLD) in order to detect the attack model. The researchers have identified five different classes of attacks and have successfully utilized KLD to identify frauds, such as multiple readings from the customer's tariff data. Additionally, in Badrinath Krishna et al. (2015), the authors proposed and utilized a Principal Component Analysis (PCA) to monitor consumption readings and detect any anomalies that might occur during the reading and billing process based on historical values. The work on O'Toole et al. (2019) was continued in Krishna et al. (2018) to deal with different signal processing-based approaches for finding irregularities in metering frauds involving Distributed Energy Resources (DERs) like wind and solar.

As the public, commercial, and academic attention is increasing toward novel use-cases of ML across various domains, new vulnerabilities are also emerging. Some vulnerabilities can affect how the output of the utilized ML algorithm will change via carefully and maliciously designed input data. Such attacks are defined as adversarial

attacks and can have devastating consequences, especially within the sectors of health and energy. In Finlayson et al. (2019), the researchers evaluated a tumor detection algorithm and its vulnerability to adversarial attacks. According to the results, after adversarial attacks, the algorithm started to classify benign tumors as malignant, which can lead to healthcare fraud. According to the authors, it is also possible to carry out a similar but reverse attack type in which benign tumors can pass for malignant ones, endangering the patient's health. Moreover, research on adversarial attacks is also occurring within the commercial domain. Albeit on a large scale. Kurakin et al. (2016) focused on adversarial training on large ML models, which are especially prone to adversarial attacks due to a large number of input parameters. According to the study results, adversarial training as a mitigation mechanism provided added robustness toward adversarial attacks. However, it has also been noted that the adversarial training model does not help against iterative adversarial attacks. Yet, iteration-based adversarial attacks are less likely to propagate across ML networks. Thus, indirect robustness is inherently provided. Since Szegedy et al. first pointed out that Deep Neural Networks (DNNs) could be attacked by adversaries in 2014 Szegedy et al. (2014), a lot of research has been done to both come up with new ways to attack adversaries and make DNNs more resistant to attacks by making their models stronger (Huang et al. 2020; Ozgur Catak et al. 2020; Qayyum et al. 2020; Sadeghi et al. 2020). The vulnerabilities inherent in deep learning models pose formidable challenges in the face of adversarial attacks, rendering them intricate to safeguard effectively. One notable vulnerability is their heightened sensitivity to minor alterations in input data, leading to unpredictable outcomes in the final output of the model. Traditionally, adversarial attack strategies predominantly center on perturbing input instances to maximize the model's loss. Over the past few years, a plethora of adversarial attack algorithms have been proposed, reflecting a concerted effort to explore and exploit the vulnerabilities of deep learning models. These algorithms seek to manipulate the model's decision boundaries, thereby inducing misclassifications or erroneous predictions.

Presently, research is deficient in a cohesive approach that combines several AI applications for the purpose of identifying anomalies and ensuring cyberphysical security. This gap is especially noticeable when there is a lack of methods that use AI algorithms to challenge and improve one another's skills. Furthermore, there is a significant deficiency in the existing body of knowledge about the dependability of AI-driven models when applied to trusted critical infrastructure. There is a lack of research specifically examining the comparison between the ability to withstand cyber-physical attacks and the effectiveness of anomaly detection. This article aims to close this gap by using a variety of AI methods to systematically improve the digital immutability and trustworthiness of the investigated critical infrastructure. The objective is to enhance critical systems' resilience, including digital and physical components, specifically emphasizing power infrastructure.

3 Background and interdisciplinary framework

This section serves to provide an appropriate theoretical foundation relevant to the proposed approach. This involves designing the framework of the cyber-physical-social system in connection to the proposed content. Moreover, it presents fundamental background about theoretical insights into cyber-physical security and the resilience of power systems.

3.1 Cyber physical and social systems

In order to provide a holistic view of smart grid applications, The Smart Grid Architecture Model (SGAM) was developed under the mandate of M/490 (Bruinenberg et al. 2012), and was particularly adopted, utilized by the European Union (EU) countries, industry, and academia (Uslar et al. 2017). However, SGAM and similar architecture models with multiple layers did not consider the enablers for deep digitalization such as Artificial Intelligence (AI), Machine Learning (ML), and Distributed Ledger Technology (DLT). Thus, a reference model with multiple layers that can provide a holistic view while considering mentioned enablers for deep digitalization was developed in order to provide insight towards next-generation smart grid systems (Cali et al. 2021) and visualized in Fig. 1.

The layers in the CPSS model for power systems are:

- Energy policy and regulatory layer: Responsible for new policy development, supervision, and management of energy legislation and regulations. Policymakers propose and develop new policies to ensure the needs of the energy industry are satisfied while considering energy security and emissions. Transmission System Operators (TSOs), Distribution System Operators (DSOs), and other market participants are legally obligated to comply with energy policy and regulations.
- Business layer: There are many stakeholders in the modern power markets such as; utilities, TSOs, DSOs, trading companies, investors, prosumers, etc. This increased segmentation and participation is due to deregulation and decentralization of the power markets and is expected to grow in the future due to The Green Digital Shift (Cali et al. 2021). The business layer is affected by the regulation policies, legislations, and economic metrics such as the Levelized Cost of Electricity (LCOE), Levelized Cost of Storage (LCOS), Net Present Value (NPV), Return on Investment (ROI), etc. There-

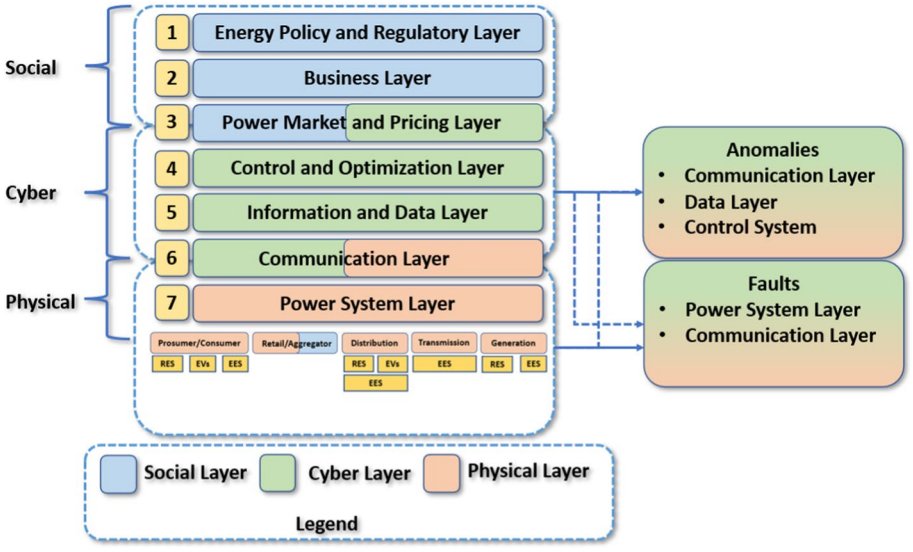

Fig. 1 CPSS model for power systems (Cali et al. 2021)

fore, comprehensive benefit and cost analysis needs to be performed diligently for any investment decision.

- Power market and pricing layer: Responsible for managing the physical power grid operations, energy and financial transactions, settlements, and data exchange (communication) between the market participants. Due to the deregulation and decentralization of the power industry, independent energy producers, non-utility producers, and prosumers were able to enter the power market to trade energy and provide ancillary services. This in return allowed the utilization of local energy markets and made the operation, and pricing of power markets an important research area.
- Control and optimization layer: The power industry adopted the Supervisory Control and Data Acquisition (SCADA) systems over the last two decades, and more recently began to utilize Phasor Measurement Units (PMUs) to monitor daily system operations, optimizations, and control of the power grid. Data analytic technologies such as AI and ML as well as DLT are being increasingly integrated into the existing power grid infrastructure for better security of supply, control, optimization, and cyber-security.
- Information and data layer: Responsible for data processing, analysis, and cyber-security aspects of the smart grids. Additionally, integration of DLT which supports smart contracts for various transactions and tokenizations, coupled with AI, ML is an active research area within this layer. The main focus of this article is mostly related to this layer.
- Communication layer: Responsible for the integration of communication protocols (Huang et al. 2021) across the different layers and one of the critical layers in modern smart grid systems due to cyber-security, reliability, scalability, and power consumption.
- Power system layer: Responsible for day-to-day operations of the physical components of the power system such as generators, transmission and distribution infrastructures, consumers, prosumers, etc.

3.2 Phasor measurement units

PMUs can be defined as devices that are able to measure positive sequence voltages, currents and calculate the phase angles and the Rate of Change of Frequency (ROCOF) with high accuracy and sampling rate. The probability of a major blackout has driven the global power industry towards implementing Wide Area Measurement Systems (WAMS) where PMUs are an essential key player since their high sampling rate, specially designed high-speed communication protocols, and time-synchronized measurements via Global Positioning System (GPS) offers to locate anomalies quickly and deploy preventive measures to avoid large faults (Phadke and Bi 2018; De La Ree et al. 2010).

Traditional measurement techniques such as Supervisory Control and Data Acquisition (SCADA), don't allow high-speed measurements. Thus, the ability of PMUs to achieve high sampling rates offers an excellent opportunity for grid operators to get clearer and more accurate information regarding the state of the grid (Ren et al. 2018). An additional benefit of PMUs is their ability for synchrophasors, which are time-synchronized phasor measurements across different locations in the grid, which can be used for providing information on both the supply and the demand side within the same timeframe. Therefore, allowing vital information to be displayed, and analyzed in a fast and accurate manner (Vicol et al. 2013).

The high amount of data that the PMUs are able to collect is considered an advantage and a big resource for the system operators. However, accurate and efficient Machine Learning (ML) algorithms that are specifically built for handling such loads are needed in order to utilize the said resources in the best way possible (Garza and Mandal 2022).

Future power systems may benefit from the PMU's capacity to synchronize each measurement across a vast region utilizing GPS. The phasors that have been estimated at a specific time stamp are referred to as synchrophasors. To ensure that the measurements took place at the precise same moment, WAMS relies on synchronizing the time stamps over a wide area. However, because the anti-aliasing filter applies a phase delay to the input signal, it interferes with synchronization. Both the frequency of the signal and the filter's properties affect this delay. Because the measurement is done after the filter, the PMU must make up for this delay for the synchronization to be accurate.

The connection to satellites in Earth's orbit ensures that the internal clock of a GPS device is extremely accurate. The GPS transmits a signal that pulses once per second to transmit this data. A sampling clock that is phase-locked to the GPS signal is used to synchronize the PMUs. The PMU generates the time stamps at a frequency that is multiplied by the nominal frequency of the power system. The analog wave patterns for each phase are digitalized using an analog-to-digital converter. Each sample is synchronized with a location and a time stamp with an accuracy of one microsecond using the GPS receiver and phase-lock oscillator. The samples are sent to a receiver at up to 60 Hz after the phasors have been time-tagged and found. The block diagram of the steps taken by the PMU is illustrated in Fig. 2

3.3 Anomalies within the power system

The anomalies within the modern power systems can be categorized into two distinct variations, one being physical and the other cyber (Halden et al. 2023), with various interactions between the two, as illustrated in Fig. 3.

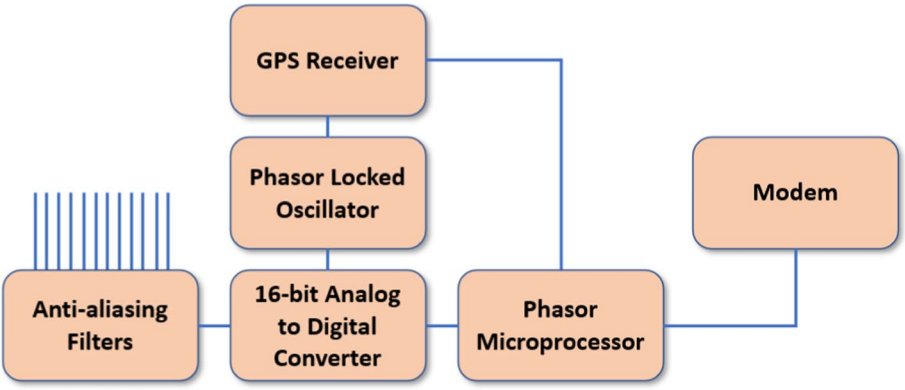

Fig. 2 Block diagram of PMU, adapted from Vicol et al. (2013)

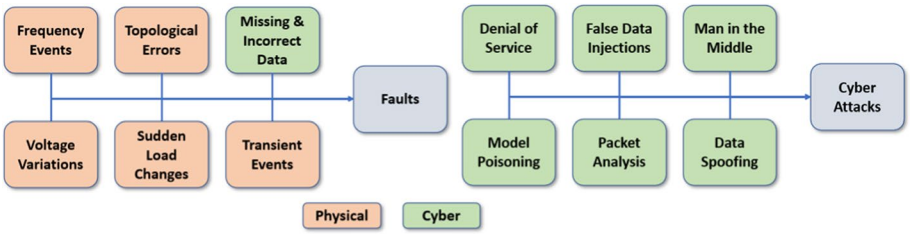

Fig. 3 Example of physical faults and cyber anomalies within the power system

3.3.1 Frequency events

Frequency-based anomalies are deemed to be critical phenomena in power systems that can have far-reaching implications regarding grid stability, reliability, and safety. Such anomalies can be characterized by a sudden and significant deviation from the nominal system frequency.

One of the most common reasons for frequency-based anomalies within the power system domain is due to sudden loss of generation or load. When a large-scale power generator or a significant load is being disconnected from the grid, it can cause an immediate drop in the overall grid frequency. Alternatively, if a large-scale generator or load suddenly gets connected to the grid, this can cause a major increment within the grid frequency. Thus, rapid detection and mitigation of such problems are essential for grid stability.

Another type of frequency-based anomaly within the power grid can be caused by physical component faults, such as short circuits or equipment failures. For such events, protective relays and circuit breakers can be employed within the grid system in order to detect and isolate the fault region. Thus, preventing a cascading event throughout the grid.

In order to detect and analyze such anomalies, power system operators rely on PMU systems to continuously monitor the grid's frequency and provide real-time data to various algorithms and automated systems for quick action taking. However as mentioned, the high sampling rate of PMUs require advanced data handling capabilities due to rapid data gathering, as well as quick and efficient algorithm designs to handle the gathered data and perform the anomaly detection before a total system collapse. Therefore, ML techniques as well as statistical data analytics are increasingly employed to process, identify and predict such faults (Yang et al. 2018; Rafferty et al. 2018).

3.3.2 Sudden load change

Sudden load changes (SLCs) occur when loads are suddenly added or removed from the power system and can affect the quality of the delivered power. The introduction and removal of loads, in ideal cases, should be done while considering load management techniques such as increasing generation during the startup of an industrial process (Styvaktakis et al. 2003). However, SLCs can also occur during anomalous situations such as a generator fault, which needs to be taken offline. Similarly, the same can happen if a circuit breaker is triggered in order to clear a fault and protect the system. Thus, various topological errors and SLCs are closely intertwined and detection and handling of SLCs are important for ensuring good Power Quality (PQ) (Pardha Saradhi et al. 2020).

Yang et al. (2018) considered sudden load changes and how high sampling frequency of the PMUs coupled with strict latency requirements can lead to additional problems for the system operators. Thus the authors proposed a fog computing framework that distributes the required computational load across different edge devices, hence increasing the anomaly detection times by reducing the network propagation rate. The researchers implemented k-NN and Singular Spectrum Analysis (SSA) algorithms across the distributed edge devices simulated under the IEEE 16 machine 68 bus system and demonstrated that fog computing can reduce the data flow End-to-End (ETE) delay by 50%.

3.3.3 Transient events

As defined by Styvaktakis et al. (2003, 1995), transient events can be defined as short events on voltage and current signals in a given power grid and can be categorized into three main sub-groups as:

- Events that happen over a long duration of time and adjust or change the voltage magnitude of the fundamental frequency. Such events have the potential to create voltage sags or swells ranging from 50 ms to several seconds.
- Events that happen over a short duration of time and change the voltage magnitude. An example of such events can be fuse-cleared or self-extinguishing faults.
- Events that the fundamental voltage magnitude is not important, such as during a lightning strike.

The authors in Zhou et al. (2016) focus on detecting both impulsive and oscillatory transient events in a distribution network by utilizing micro PMUs. The authors utilized kPCA algorithm for binary decision-making combined with a pSVM to distinguish event types while considering both labeled and unlabeled data information. According to the results, the proposed model has an accuracy of over 93% and can be used to detect anomalies that can occur due to transient events in the distribution network.

Impulsive transient Impulsive transients are phenomena where a sudden change occurs in the steady state condition of voltage, current, or both and are generally associated with lightning strikes due to them being the most common cause. An example of an impulse transient event is illustrated in Fig. 4

Oscillatory transient Oscillatory transient events denote the rapid changing of the voltage or current values and can be classified according to their frequency rate, as demonstrated in Table 1.

Table 1 Categorizing of transients (1995)

Categories		Spectrum	Typical duration	Typical magnitude
Impulsive	Nanosecond	5 ns rise	< 50 ns	
	Microsecond	1 µs rise	50 ns to 1 ms	
	Millisecond	0.1 ms rise	> 1 ms	
Oscillatory	Low frequency	< 50 kHz	0.3–50 ms	0–4 pu
	Medium frequency	5–500 kHz	20 µs	0–8 pu
	High frequency	0.5–50 MHz	5 µs	0–4 pu

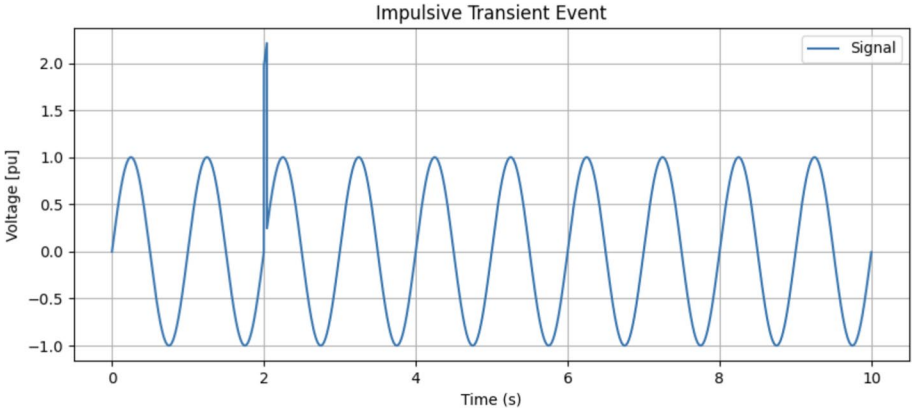

Fig. 4 Impulsive transient event occurring in a power system, adapted from

The majority of the high-frequency oscillatory transient events are a result of some type of switching event happening in the grid, usually as a response to an impulsive transient event such as a lightning strike. Meanwhile, medium frequency oscillatory transients can be the result of back-to-back capacitor energization, which happens when a capacitor bank gets energized next to an already in-use capacitor bank.

Oscillatory transients with low-frequency values can have various causes, similarly to the cause of medium-frequency oscillatory events, some may be caused due to capacitor bank energization, which can induce transients between 300 and 900 Hz, while ferroresonance and energization of transformers can result in transients below 300 Hz, making them contained mostly on sub-transmission and distribution system levels.

3.3.4 Topological errors

Line status error and substation configuration error are the two main types of power system network topology faults that are caused by inaccurately reported circuit breaker status (Abur and Exposito 2004). The former refers to incorrectly excluding or including transmission lines from the network model, while the latter refers to bus splitting or merging errors at the substation level. The summary of these errors can be seen in Fig. 5 illustrated within a 2-bus system.

In El Chamie et al. (2018) the authors proposed an anomaly detection technique for power grids that builds machine learning models with physics-based features using data from PMUs. Instead of using the conventional steady-state anomaly detection algorithms, the resulting model finds anomalies based on their transient features. The placement of the anomaly detection algorithm on the distribution grid allows for quicker anomaly identification and better localization. The results of simulations were performed on the IEEE 34-node feeder and demonstrate that the anomaly detection algorithm performed better to detect various classes of anomalies such as single line to ground faults.

Arefin et al. (2022) focuses on detecting topological errors and islandings within the power network by utilizing PMU data. The researchers specifically utilized frequency and phase angle data coupled with time series anomaly detection techniques to identify and detect the islanding events. The results of the study can help to provide more symmetrical

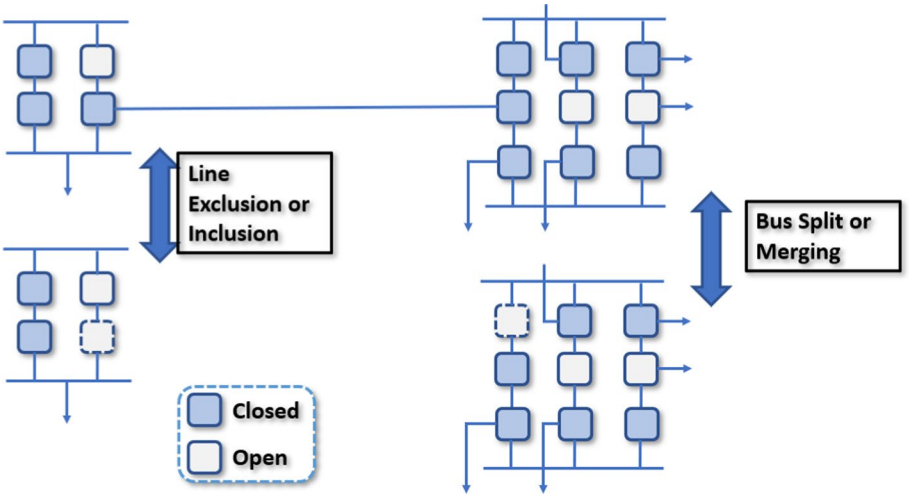

Fig. 5 Topological error illustration in 2 bus system, adapted from Choi and Xie (2017)

and improved PMU data analytics and better islanding event detection for enhanced grid reliability.

3.3.5 Voltage variations

Voltage variation-based faults pose a critical concern within the power systems due to their potential for wide-scale disruption if left unchecked. Such faults denote a range of voltage deviation, which can include over-voltage and under-voltage conditions. Such anomalous conditions can occur due to various conditions, including:

- Transient faults: Such faults can be the result of natural phenomena, such as lightning strikes, or operational reasons such as sudden switching events, as well as cyber-attacks. If not mitigated in a quick manner, these transient faults can propagate through the system and affect a widespread area,
- Physical equipment fault: Such faults occur when a physical component of the power grid, such as transformers and circuit breakers stops functioning properly, which can lead to over or under-voltage events. For example, a transformer fault can lead to a significant voltage drop within the surrounding area where the transformer is being located. Thus, impacting consumers and prosumers alike across the distribution grid,
- Sudden load changes: Rapid fluctuations within the load demand, such as during heavy motor startup or other heavy industrial processes can result in voltage sags or swells, which can be especially problematic for sensitive equipment, as in the case of health equipment within hospitals.

Voltage variation anomalies can have major consequences for both utilities and end-users, where equipment damage is one of the notable outcomes. Thus, requiring expensive repairs or total replacements. Additionally, within industrial-scale applications, voltage variation faults can disrupt production processes. Therefore, cause downtime and financial loss. In order to mitigate such faults, power system operators can utilize:

- Voltage regulation: Automatic Voltage Regulators (AVRs) and tap-changing transformers can be employed by the grid operators in order to maintain the voltage levels within an acceptable range. Thus, ensuring a stable and reliable power supply to the end customers,
- Fault detection and isolation: Advanced monitoring systems and tools such as PMUs can detect voltage sags or swells within a quick timeframe and allow the grid operators to isolate the affected area in order to minimize disruption within the overall grid,
- Transient voltage suppressors: If the cause of the voltage variation is detected to be due to natural phenomena such as lightning strikes or equipment error, surge arresters, and other protective equipment can be deployed in order to reduce the transient voltage spikes.
- Load management: Load shedding and shifting is another tool the grid operators can utilize in order to manage the voltage fluctuations within the power system. Reducing the likelihood of voltage variations during peak demand periods.

3.3.6 Missing and incorrect data

It is inevitable that faults will occur seldomly while processing a large number of data points such as PMUs (Karpilow et al. 2020). The term "Bad Data" can include missing data where a problem occurred during measurement recording or data that is unrepresentative of the real situation of the analyzed power system. There can be a number of reasons for bad data quality such as equipment failure (sensor errors), communication problems between the devices or a combination of both.

Since PMUs record and handle high volumes of data, a short burst of error during data transmission can result in high amounts of missing data points. However, since PMU measurements are time synchronized, the missing points can actively merge back into the dataset. This, however, will lead to problems for real-time SE or anomaly detection in the system, where continuous data feed is required.

An example of bad data can be seen in Fig. 6 where four different types of bad data are illustrated (Tinawi 2019). As can be seen, sensor malfunctions can lead to major oscillations or noise which is higher than the original. Additionally, the same malfunctions can result in measurement spikes either in the form of a single data point or over a time frame. Meanwhile, synchronization error measurement drifts, where over time the errors might add up and result in measurements being even less representative. Finally, malfunctioning measurements can lead to either high or low-magnitude offsets in the data.

In Amutha et al. (2021), the authors used the density estimation technique, which is based on the Gaussian Mixture Model, to take into account all the features and identify anomalies in real-time streaming PMU multi-variate data in a smart grid. The distribution

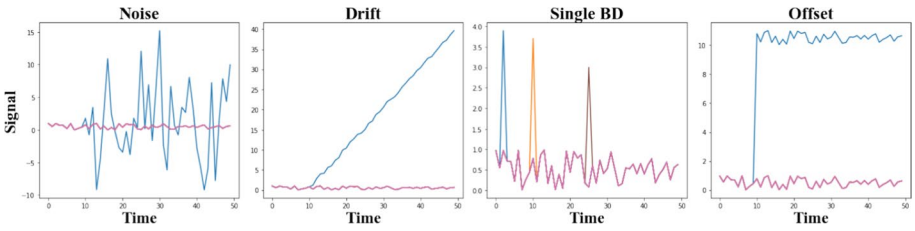

Fig. 6 Different types of bad data examples (Karpilow et al. 2020)

of normal PMU data followed a GMM with five mixtures, and by using principal component analysis, the pertinent features were chosen in six different combinations. The accuracy of classification increases with the number of features used, but it is impacted by random initialization, which makes it harder to determine whether a set of data is normal or anomalous. The system is able to detect anomalies within the selected window in the streaming PMU data with a low false positive rate and an F1 score of 1 for the chosen features, according to the performance evaluation of testing data with 13 blocks in online mode for 2, 3, 8, 10, 12 and 16 features. The suggested framework was tested with streaming data in both online and offline modes, and the results of the experiments show that the proposed methodology can perform anomaly detection.

Deng et al. (2020) considered four types of anomalies that can happen during PMU data retrieval as erroneous patterns, high-frequency interferences, missing points, and random spikes in the data. A deep learning CNN model was employed for real-time anomaly detection in the Jiangsu power grid located in China and was able to perform anomaly detection with a continuous data stream, hence being computationally efficient for easy implementation. The proposed CNN model achieved 97.71% accuracy over the testing data and can effectively detect data-based anomalies within synchrophasor measurements.

3.3.7 Denial of service attacks

The FDI attacks are not intended to be detected, meanwhile for Denial of Service (DoS) or Distributed Denial of Service (DDoS) attacks will be immediately detected and start to affect the overall system. This type of attack is constructed to overload the utilized communication network between data transmitting devices and prevent the normal data flow from occurring (Ramasubramanian et al. 2022). Thus, DoS or DDoS detection is not based on detecting subtle differences in the measured data, but on understanding that an attack is ongoing instead of basic communication error within the measurement devices.

3.3.8 False data injections

False data injection (FDI) attacks work by getting unauthorized access to the data stream and altering the retrieved measurements between the measuring and data collection points (Risbud et al. 2019). FDI-type attacks can have massive effects on the stable operation of the power grid, as the measured data is constantly in use for SEs and production, load balancing. Thus, FDIs can lead to rapid destabilization of the whole system with possible catastrophic consequences.

In the case of microgrids, a common way to perform destabilization is via GPS spoofing and PMU data alterations (Risbud et al. 2019), which is possible due to PMUs being time synchronized. The actual spoofing is done by achieving access to the data stream and sending intentionally fabricated measurements to the Phasor Data Concentrators (PDCs) to inject false data, change the timestamps of the measurements and lead to voltage/current magnitudes and angles to be unsynchronized (Jafarnia-Jahromi et al. 2012).

The authors in Pal and Sikdar (2014) perform a Gauss-Newton iterative method and obtain the transmission line parameters while also reducing the residuals. Prior to the analysis, the authors assumed the nominal values of the line parameters to be known. After performing a chi-square test based on the known and estimated line parameters which were modeled as a state estimation problem, possible anomaly detection scenarios were performed. According to the results, the authors were able to detect anomalies

that can arise due to FDIs in an example topology and claim the same methodology can be scaled up to the whole power system with minimal needs.

Another FDI detection work was performed by Wang et al. (2018) by utilizing distributed edge devices and deep autoencoders which is a generative deep learning model. By feeding the newly acquired data into the deep autoencoder and computing the reconstruction errors after training, the anomaly measurer may be used to evaluate the likelihood that the FDI exists. The suggested framework can be used to detect FDI in the entire power system by merging the local results received from various anomaly measurers and working with other information sources. Additionally, to prevent false positives, a delayed alert triggering algorithm was also implemented, which also benefits towards an improved noise immunity.

3.3.9 Man in the middle attacks

A Man in the Middle (MITM) attack is a cyberattack in which the attacker places oneself between two parties who believe they are directly communicating with each other without any third-party involvement, and then secretly transmits or modifies their messages. Active eavesdropping is one type of MITM attack in which the attacker establishes separate connections with the victims and relays messages between them to give the impression that they are speaking directly to one another over a private connection when in reality the attacker is in control of the entire conversation (Sivasankari and Kamalakkannan 2022). All essential messages sent between the two victims must be intercepted by the attacker, who must then introduce fresh and malicious messages. In many cases, this is simple; for instance, a person within the communication range of an unsecured Wi-Fi access point could act as a man-in-the-middle attacker. The wormhole attack is another example of a sub-MITM assault in which the attacker penetrates the network and listens to network activity without changing any of the original communications between the conversing parties. Meanwhile, Sinkhole Man in the Middle Attacks (SMITM) can crash the entire network connection by generating a large volume of network traffic via sending requests and routing information to nearby nodes while also broadcasting falsified information.

3.3.10 Data spoofing

Cyberattacks that use spoofing often take advantage of established connections by pretending to be someone or something that the victim is familiar with. These messages may even be tailored to the victim in some situations, such as whale phishing attacks that use email spoofing or website spoofing, to persuade the victim that the contact is genuine. A user is more likely to be the victim of a spoofing attack if they are not aware that communications might be falsified.

A successful spoofing attempt could have catastrophic consequences. Sensitive personal or business data may be stolen, credentials may be gathered for use in fraud or future attacks, malware may be transmitted via malicious links or attachments, trust relationships may be used to gain unwanted network access, and access limits may be disregarded. They might even conduct a MITM or DoS/DDoS attack, or malicious code injections into the system.

3.3.11 Package analysis

The power grid is a critical component of the modern society. Thus, rendering it a primary target for cyberattacks. Among the diverse array of available cyberattacks, package analysis has emerged as a prominent concern within the power system domain (Tu et al. 2018).

Data package analysis-based attacks denote the interception of data packages that are being exchanged by various IoT equipment within the grid network. Such attacks can be utilized by malicious entities in order to gain valuable insight into the communication patterns of the infrastructure, power grid vulnerabilities, and various other sensitive information. Therefore, the objective of such attacks can be summarized as:

- Information gathering: Malicious actors can seek to gather information regarding the power system architecture, utilize communication protocols and procedures in order to leverage weaknesses within the power system,
- Vulnerability identification: Through the network analysis, attackers can try to pinpoint vulnerabilities within the power system and open a way for future and more critical attacks in order to disrupt the power grid,
- Cyber-physical attacks: Data package analysis can also serve as a precursor for cyber-physical attacks. Where information regarding the physical components of the grid can be gathered for physical sabotage.

In order to mitigate the risk of data package analysis within the power grid, a multi-strategy must be utilized. Such strategies include state-of-the-art encryption protocols that can mitigate eavesdropping and Intrusion Detection Systems (IDSs), enhancing the power system's capabilities by providing real-time intrusion alerts. However, it should be noted that as technology keeps evolving, the techniques that are being employed by malicious actors evolve in tandem. Therefore, power utilities as well as policy actors need to remain vigilant for emerging threats.

3.3.12 Model poisoning

As ML started to be used within the cyber defense industry, model poisoning attacks started to evolve together with the defense algorithms, much like in a game of cat and mouse. Hence, the first examples of model poisoning attacks against ML systems were focused on evading spam e-mail classifiers.

ML model poisoning attacks happen when the attacker can and will inject specially constructed bad data into the ML model training dataset, resulting in the model learning something it shouldn't. As the attack is done to the training dataset, the most common result is that the models' decision-making boundary shifts in such a way, in the case of anomaly detection in PMU systems, this will show itself as actual anomalies within the power system being categorized as normal operation conditions or vice versa.

Model poisoning can happen in two ways, the ones that target the ML models' availability and the ones that target its integrity, which is also known as backdoor attacks. The initial model poisoning attacks were the first type, which aimed to inject specially crafted bad data into the training pool in order to shift the models' boundaries, making it practically useless. The newer type of attack is the backdoor attacks, which are much more sophisticated compared to the availability type attacks and aim to keep the ML model as intact as

possible with the exception of adding a backdoor to the system. In this context, the backdoor can be defined as a type of input of which the model owner is not aware, however, the attacker party can utilize it to get the ML system to do what they want, such as classifying faulty operations as operations under normal conditions, by using a special key attached to the input data, so the ML system can classify it automatically as normal operation due to installed backdoor.

In Roy et al. (2020), the researchers performed availability type model poisoning to a synthetic PMU dataset where three attack strategies were carried out as step attacks where the current values in poisoned data used for training purposes increase by an average value during the whole training time period, ramp attack where the current values used during the training phase increases up to a certain value and then decreases again to starting point and finally the mirroring type, where the snapshot of current time series dataset was used over and over during the training phase in order to falsify the model. According to the findings of the research, the harmonic to the arithmetic mean ratios of the power system is a stable and effective way to determine if there were any model poisoning, as even if the attacker has knowledge about the time series data used during the testing, the defenders can identify anomalies happening in the power system in real-time with a 91% accuracy.

Meanwhile, Bhattacharjee et al. (2022) developed an anomaly detection algorithm based on the Ordinary Least Squares (OLS) regression model which focuses on microgrids and thus, utilized smart meter data instead of PMU dataset. According to the performed research, using L1 norm instead of L2 norm helps to protect the ML model against model poisoning, as the L1 norm has a gradient gradient-shattering effect which does not allow for the calculation of accurate gradients during the training phase. Hence, limits the attackers' ability for gradient shifting.

3.4 Artificial intelligence and machine learning

The use of AI/ML has gained significant attention in the domain of power systems as well as in several other industries. This section presents appropriate AI/ML techniques for the proposed approach.

3.4.1 Long short term memory

Long Short Term Memory (LSTM) networks replace hidden units within the Recurrent Neural Networks (RNNs) by *memory cells* which is constructed via three gates (Hochreiter and Schmidhuber 1997). In addition to the utilization of said gates, the LSTM networks also include a cell state vector, which can be denoted as $\mathbf{C}^t$ in order to keep track of the critical information within the network. The structure of an LSTM network is illustrated in Fig. 7. For each time step during the modeling, the information can either be added or removed via the input and forget gates, respectively. Meanwhile, the output gate is utilized for deciding which information to keep for the next hidden state and the output.

The initial step within an LSTM network is for the cell to decide which information from the previous hidden state (denoted as $\mathbf{h}^{(t-1)}$) and from the input (denoted as $\mathbf{x}^t$) is surplus information and hence should be forgotten. This step is performed by multiplication of the concatenate of $\mathbf{h}^{(t-1)}$ and $\mathbf{x}^t$ via a weight matrix (denoted as $\mathbf{U}$). Additionally, similar to Multi Layer Perceptron (MLP) and RNNs, a bias (denoted as $\mathbf{b_f}$) is added via a sigmoid function, as mathematically shown in Eq. (1a).

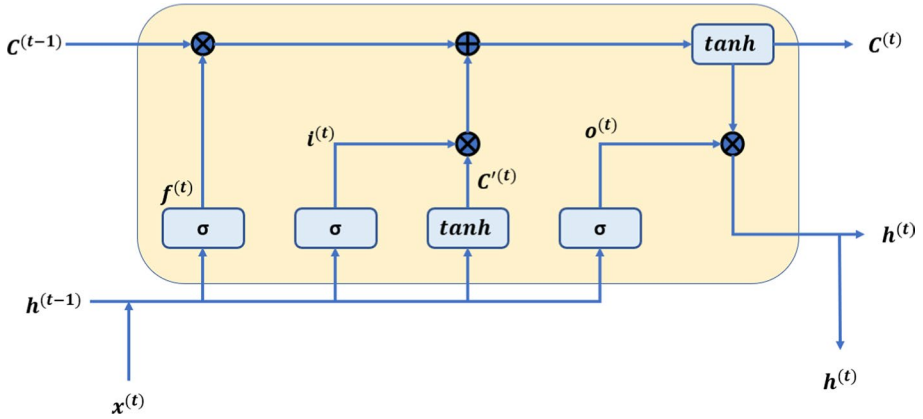

Fig. 7 LSTM Cell (Henriksen et al. 2022)

Meanwhile, Eq. (1b) helps to determine which information should be kept within the cell by updating the current cell state. In order to achieve this, a candidate vector (denoted as $\tilde{\mathbf{C}}$) is utilized, as shown in Eq. (1c) via the help of a candidate activation function, namely tanh in order to keep the values within −1 and 1. In order to calculate the updated cell states, the *unimportant* parts need to be forgotten, which is achieved by multiplying the previous cell state with the forget vector. After the forget gate, the new information can be added by multiplying the input vector with the candidate vector, as shown mathematically in (1d). As the last step, in order to create a hidden state, an output vector is created as shown in Eq. (1e) where $\mathbf{W}$ denotes the output weights and $\mathbf{b_o}$ the output bias (Henriksen et al. 2022).

$$\mathbf{f}^{(t)} = \sigma(\mathbf{U} \cdot [\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}] + \mathbf{b_f}), \tag{1a}$$

$$\mathbf{i}^{(t)} = \sigma(\mathbf{V} \cdot [\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}] + \mathbf{b_i}), \tag{1b}$$

$$\tilde{\mathbf{C}}^{(t)} = tanh(\mathbf{W_c} \cdot [\mathbf{h}^{[}(t-1), \mathbf{x}^{(t)}] + \mathbf{b_C}, \tag{1c}$$

$$\mathbf{C}^{(t)} = \mathbf{f}^{(t)} * \mathbf{C}^{(t-1)} + \mathbf{i}^{(t)} * \tilde{\mathbf{C}}^{(t)}, \tag{1d}$$

$$\mathbf{o}^{(t)} = \sigma(\mathbf{W} \cdot [\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}] + \mathbf{b_o}), \tag{1e}$$

$$\mathbf{h}^{(t)} = \mathbf{o}^{(t)} * tanh(C^{(t)}). \tag{1f}$$

3.4.2 Convolutional neural networks

Convolutional neural networks are particularly adept at recognizing patterns. As a result, CNN has proved to be extremely useful in image recognition due to the way it breaks the image down into its various components, making it simpler to recognize unique elements

in an image. Different convolutional layers are used to accomplish this by *looking* for various patterns.

For images, a 2-dimensional search grid is used. However, as time series data is composed of different patterns transpiring within different frequencies, a 1-dimensional search grid can be utilized, as well as a classical 2-dimensional grid if there are multiple time steps existing simultaneously within the dataset (Goodfellow et al. 2016). The mathematical formulation of a convolution can be seen in Eq. (2), which, as the name implies, is heavily utilized in CNNs. Unlike MLPs, CNNs do not need to be fully interconnected, hence can utilize *sparse interactions*, which aids in simplification of input data. Compared to MLPs, another advantage of CNN can be noted as the parameter sharing for future maps, which in combination with sparse interactions, helps to achieve lesser computational times and memory usage.

$$(f * g)(t) = \int_{-\inf}^{\inf} f(\tau)g(t-\tau)d\tau \tag{2}$$

The convolutional layers can be divided into three main stages as: convolution, non-linearity and finally, pooling. The convolution step is responsible for performing multiple convolutions in order to transform the input into various sets of outputs, which is often called a feature map. For the general convolution equation, as stated in Eq. (2), f will denote the input tensor, whereas the g is the kernel, which is itself another tensor for feature extraction. In the second step, a non-linearity is added as the first convolutional step is linear. In the final step, the input is divided into various and equal-sized rectangles, where the size is equal to kernel size, which is then simplified. For the simplification process, max pooling is one of the most heavily used techniques (Goodfellow et al. 2016), which works by taking the maximum value in a given area.

3.4.3 Adversarial machine learning

Adversarial machine learning is a field that focuses on studying the vulnerabilities of AI models to intentional attacks and developing robust defenses against such attacks. Adversarial attacks aim to exploit the weaknesses in AI algorithms by intentionally manipulating input data to mislead or deceive the models' predictions.

In the context of anomaly detection algorithms in energy systems, adversarial attacks can undermine the effectiveness of these algorithms by introducing subtle perturbations or crafting malicious inputs. Adversarial attacks against anomaly detection can be categorized into two main types: evasion attacks and poisoning attacks.

Evasion attacks, also known as adversarial perturbations or adversarial examples, involve manipulating input data to make anomalies appear normal or to hide anomalies from detection. These attacks aim to evade the anomaly detection algorithm by perturbing the data in ways that are invisible to human observers but can mislead the AI model's predictions.

On the other hand, poisoning attacks involve injecting malicious or deceptive data during the training phase of the anomaly detection algorithm. By poisoning the training data, adversaries can manipulate the AI model's learned patterns and decision boundaries, leading to compromised performance during anomaly detection.

Adversarial machine learning techniques, such as adversarial training and defensive distillation, have been proposed to enhance the robustness of AI models against adversarial attacks. These techniques involve augmenting the training process with adversarial examples

or introducing additional defenses to detect and mitigate adversarial manipulations in the input data.

Understanding the vulnerabilities introduced by adversarial attacks and developing effective defense mechanisms are essential for ensuring the reliability and security of anomaly detection algorithms in energy systems.

In the following sections, we will delve deeper into the cybersecurity challenges related to anomaly detection algorithms in energy systems and discuss the specific adversarial attacks and defense strategies relevant to this context.

Basic iterative method (BIM) The Basic Iterative Method (BIM) is a popular iterative attack technique in adversarial machine learning. It aims to generate adversarial examples by perturbing the input data in small steps while ensuring that the perturbations stay within a specified epsilon (ϵ) boundary.

The BIM attack starts with an initial adversarial example $\mathbf{x}^{(0)}$, which is typically a slightly perturbed version of the original input example $\mathbf{x}$. Then, for a predefined number of iterations T, the algorithm computes the gradient of the loss function with respect to the input data and updates the adversarial example accordingly. The update rule for each iteration t is given by:

$$\mathbf{x}^{(t+1)} = \text{clip}_{\epsilon}\left(\mathbf{x}^{(t)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}}))\right) \tag{3}$$

Here, $\nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}})$ represents the gradient of the loss function J with respect to the input data $\mathbf{x}^{(t)}$, where y_{true} is the true label of the original input example. The term α controls the step size of the perturbations, and clip_{ϵ} is a function that clips the perturbed example to ensure that the perturbations remain within the ϵ boundary.

The BIM attack iteratively adjusts the adversarial example to maximize the loss function, aiming to fool the target model into making incorrect predictions on the perturbed input.

Momentum iterative method (MIM) The momentum iterative method (MIM) is an extension of the BIM attack that introduces momentum to accelerate the convergence toward adversarial examples. The inclusion of momentum helps overcome the oscillations often observed in the BIM attack and can result in more effective adversarial perturbations.

In the MIM attack, the update rule for each iteration t is given by:

$$\mathbf{r}^{(t+1)} = \mu \cdot \mathbf{r}^{(t)} + \|\nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}})\|_1 \cdot \nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}}) \tag{4}$$

$$\mathbf{x}^{(t+1)} = \text{clip}_{\epsilon}\left(\mathbf{x}^{(t)} + \alpha \cdot \text{sign}(\mathbf{r}^{(t+1)})\right) \tag{5}$$

Here, $\mathbf{r}^{(t)}$ represents the momentum term, which accumulates the gradients of previous iterations.

Projected gradient descent (PGD) The Projected gradient descent (PGD) attack is an iterative optimization-based method to generate adversarial examples. It aims to find the perturbation that maximizes the loss function while ensuring that the perturbed example remains within a specified epsilon (ϵ) boundary.

The PGD attack iteratively updates the adversarial example by taking small steps in the direction that maximizes the loss function while projecting the perturbed example back into the epsilon ball at each iteration to satisfy the constraint. The update rule for each iteration t is given by:

$$\mathbf{x}^{(t+1)} = \text{clip}_{\epsilon}\left(\mathbf{x}^{(t)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}}))\right) \tag{6}$$

Here, $\nabla_{\mathbf{x}} J(\mathbf{x}^{(t)}, y_{\text{true}})$ represents the gradient of the loss function J with respect to the input data $\mathbf{x}^{(t)}$, and y_{true} is the true label of the original input example. The term α controls the step size of the perturbations, and clip_{ϵ} is a function that clips the perturbed example to ensure that the perturbations remain within the ϵ boundary.

The PGD attack performs multiple iterations of the update rule to iteratively refine the adversarial example until convergence or until a predefined number of iterations is reached.

Madry attack The Madry attack, also known as the Projected Gradient Descent with random starts (PGD-RS), is a variant of the PGD attack designed to find solid, robust adversarial examples against various defences.

In the Madry attack, multiple random initialisations are used instead of starting the optimisation from a single initial adversarial example. Let i index the random initialization, and $\mathbf{x}_i^{(0)}$ represent the initial perturbed example for the i-th initialization. The attack then performs PGD iterations on each randomly initialised adversarial example. The update rule for each iteration t is given by:

$$\mathbf{x}_i^{(t+1)} = \text{clip}_{\epsilon}\left(\mathbf{x}_i^{(t)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}_i^{(t)}, y_{\text{true}}))\right) \tag{7}$$

Here, $\nabla_{\mathbf{x}} J(\mathbf{x}_i^{(t)}, y_{\text{true}})$ represents the gradient of the loss function J with respect to the input data $\mathbf{x}_i^{(t)}$, where y_{true} is the true label of the original input example. The term α controls the step size of the perturbations, and clip_{ϵ} is a function that clips the perturbed example to ensure that the perturbations remain within the ϵ boundary.

The Madry attack performs multiple iterations of the update rule for each random initialisation i to iteratively refine the adversarial examples. The final adversarial example is then selected based on the highest loss obtained across all random initialisations, making it robust against various defenses and models.

3.5 Defensive distillation-based mitigation method

The concept of knowledge distillation, initially introduced by Hinton et al. (2015), offers a means to transfer the expertise of an extensive, densely connected neural network (referred to as the *teacher*) into a smaller, sparsely connected neural network (referred to as the *student*). This approach enabled the student network to achieve performance levels akin to those of the teacher network. The original application of knowledge distillation primarily revolved around solving classification problems, a framework often called the "teacher–student" model.

Building upon this foundation, Papernot et al. (2016) extended the utility of knowledge distillation by applying it to adversarial machine learning defence. Their work showcased the technique's capacity to enhance model robustness against negative examples. The key innovation here was the introduction of knowledge distillation for the specific purpose of bolstering machine learning models against adversarial attacks.

Defensive distillation, as a machine learning framework, is primarily employed to fortify the resilience of models in classification tasks. The first step involves training the teacher model using a high-temperature parameter (T), which serves to soften the softmax probability outputs of the deep learning model. Mathematically, this process is defined as:

$$p_{softmax}(z, T) = \frac{e^{z/T}}{\sum_{i=1}^{n} e^{z_{(i)}/T}} \tag{8}$$

In this equation, n corresponds to the number of labels, and z represents the output of the final layer of the deep learning model, where $z = \mathbf{W}_n \cdot \mathbf{a}_{n-1} + b_n$. Here, $\mathbf{W}_n$ signifies the weight matrix, and $\mathbf{a}_{n-1}$ denotes the activation of the last layer.

In the subsequent step, the student model is trained using the softmax probability outputs from the teacher model but with a lower temperature parameter. The objective function for this phase is defined as:

$$\begin{aligned}\mathcal{L}_{student}(T) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\mathbf{y}_{ij}\cdot\log p_{softmax}(z_{ij},T)\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\mathbf{y}_{ij}\cdot\log\frac{e^{z_{ij}/T}}{\sum_{i=1}^{n}e^{z_{ij}/T}}\end{aligned} \tag{9}$$

In this equation, N represents the number of training samples, $\mathbf{y}_{ij}$ stands for the training label, and z_{ij} corresponds to the logit. The objective function for training the teacher model can be defined as:

$$\mathcal{L}_{teacher}(T) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\mathbf{y}_{ij}\cdot\log\frac{e^{z_{ij}/T}}{\sum_{i=1}^{n}e^{z_{ij}/T}} \tag{10}$$

Deep learning approaches have demonstrated exceptional performance in various computer vision tasks, such as image classification, object detection, action recognition, scene segmentation, and image generation. However, deep neural networks (DNNs) often require substantial training data, which may only sometimes be readily available for new tasks or domains. Several knowledge distillation methods have been proposed to train a smaller student network to emulate the predictions of a more extensive and accurate teacher network to address this issue.

Distillation techniques have also found applications in intelligent systems, including knowledge-based and rule-based systems, where the goal is to reduce the system's size and enhance its performance by improving the quality of the system's knowledge. The differences between the teacher and student models can be a form of regularization, preventing overfitting. Algorithm 1 presents the pseudocode for the distillation process.

Algorithm 1 Pseudocode of distillation

Input: Dataset D, teacher model T, student model S, loss function $\mathcal{L}$, learning rate η, number of epochs E
Output: Trained student model S
Initialize the weights of the student model S
for $e = 1$ **to** E **do**
  Randomly shuffle the dataset D
  for $i = 1$ **to** $|D|$ **do**
    Extract the i^{th} sample (x_i, y_i) from D
    Forward propagate the sample x_i through the teacher model T to obtain the output probabilities $\hat{y}_i$
    Compute the loss $\mathcal{L}$ using the output probabilities $\hat{y}_i$
    Backpropagate the loss $\mathcal{L}$ through the student model S
    Update the weights of the student model S using the learning rate η
  end for
end for
return Trained student model S

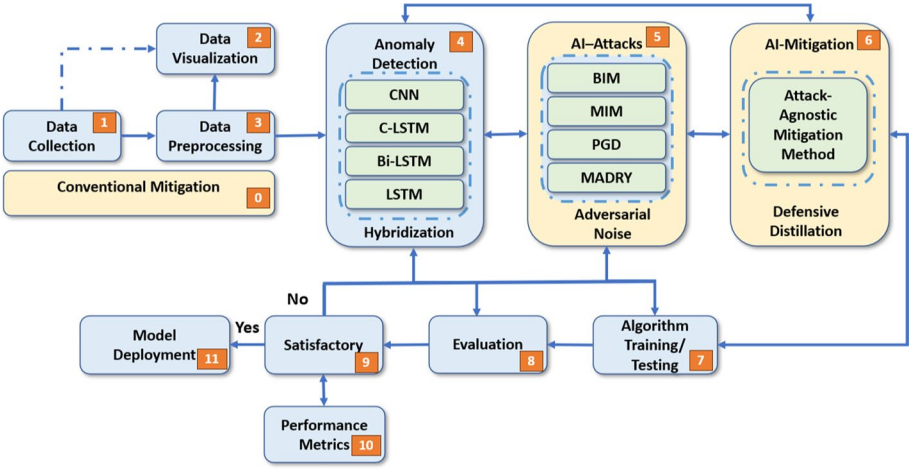

Fig. 8 The overview of the performed research methodology

4 Methodology

The flowchart depicted in Fig. 8 is a high-level representation of the integrated methodology designed to enhance the robustness of cyber-physical power systems against adversarial AI threats. This study assumes that conventional cybersecurity mitigation methods are deployed as initial phase of the framework before more sophisticated AI-based cyber-physical attacks are analyzed for the rest of the investigation. This section is dedicated to explaining the details of proposed methodology. The proposed approach was carefully designed as a multi-stage process to facilitate a thorough analysis and generate valuable insights, as well as to provide a transparent framework for readers to comprehend the processes underlying the study.

4.1 Data collection

The dataset used in this study was collected from a power grid equipped with PMUs for real-time monitoring located in Norway. Due to confidentiality resasons some details of the dataset and position of the PMU cannot be given publicly. The PMUs provide high-frequency voltage and current phasor measurements, enabling precise tracking of the power system dynamics of the Norwegian power grid where they are located. The dataset consisted of historical PMU measurements spanning one year, including data from multiple substations and transmission lines within the grid.

The safeguarding of data and the maintenance of privacy are crucial factors that must be taken into account from the initial stages of any data-intensive and data-driven solution, such as the case being presented. Stringent measures were implemented to safeguard the security and privacy of power grid data, given its sensitive nature at also earlier stages of the operations like preprocessing steps. To ensure the integrity of the data, appropriate measures were taken to maintain the cleanliness and isolation of the raw data, thereby mitigating the risk of any potential data contamination. The data handling and processing operations shall be carried out within a secure environment, implementing stringent

access controls and encryption protocols. To safeguard the confidentiality of the regional power grid, the identification of the substations and transmission lines is recommended to be anonymized in real operations.

It should be noted that this study adhered to ethical guidelines for data privacy and security. The PMU measurements used in the analysis were anonymized and aggregated to ensure the confidentiality of sensitive information. The research team obtained the necessary permissions and approvals for data access and complied with all relevant data protection regulations.

4.2 Data visualization

In order to enhance the understanding of the utilized data, as well as to act as a guide for future preprocessing steps and increase the model performance and interoperability, a manual data visualization was carried out. As a part of the data visualization, the NaN values as well as other possible data outliers were tried to be detected with the help of histograms in order to make future informed decisions. Moreover, the pre-feature selection step was carried out by assessing the relevance and correlation of variables.

4.3 Data preprocessing

The process of data preprocessing is an essential stage in energy analysis and other AI-based tasks. It involves transforming raw data into a machine-understandable format to be used as input for the AI models. Prior to commencing the data analysis procedure, it was imperative to perform comprehensive preprocessing on the gathered dataset. The initial stage of data processing encompassed a range of procedures with the objective of cleansing and converting unprocessed data into a structure that is amenable to systematic evaluation. The previously mentioned process played a pivotal role in guaranteeing the accuracy and reliability of any insights and conclusions derived from the data.

The process of data cleaning involves identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to ensure its quality and reliability. The data was initially subjected to a screening process in order to identify and address any potential anomalies or errors. For example, any instances of missing values in the PMU measurements were identified and appropriately addressed. One potential approach to address missing values is through the utilization of imputation techniques, which involve the estimation and substitution of missing values using information from other pertinent data points. In addition, the presence of outliers was identified and subsequently dealt with. Outliers have the potential to indicate measurement errors or exceptional events, and if not appropriately addressed, they can have a substantial impact on the outcomes of the analysis.

Data transformation is a process in which raw data is converted into a more suitable format for analysis or presentation. This process involves manipulating. The PMU data that was gathered exhibited a time-series structure, necessitating the implementation of transformation procedures to render it amenable for analysis. This could include various operations, including resampling, normalizing, and smoothing. As an illustration, the high-frequency data from the PMUs could be downsampled to a lower frequency, if deemed necessary for the analysis. Normalization is used as a means of standardizing measurements to a uniform scale, a practice that holds particular significance when amalgamating data from various substations or other parts of regular operations.

The process of feature extraction involves identifying and selecting relevant characteristics or attributes from raw data. The preprocessed data is utilized to extract pertinent features, which are subsequently subjected to further analysis. When dealing with PMU data, the process may entail the extraction of data pertaining to the amplitude and phase of voltage and current phasors, along with the computation of related quantities such as power and impedance.

Data partitioning refers to the process of dividing a large dataset into smaller, more manageable parts. This technique is commonly used in database management systems The data that shall undergo preprocessing steps is divided into separate datasets for training, validation, and testing purposes. The training dataset is utilized for the purpose of constructing the analytical models, while the validation dataset serves the purpose of fine-tuning the models and mitigating the risk of overfitting. Lastly, the test dataset shall be employed to assess the performance of the final model.

High-quality data establishes the foundation of anomaly detection models. This study will look at PMU data from the Norwegian TSO, Statnett, and the Texas Synchrophasor Network (TSN). The data sets consist of a total of over 50 million data points. The PMUs measure voltage magnitude, angle, power flow, and frequency. This data must be processed for further model training, prediction and anomaly detection. Data preprocessing involves removing missing values, noise filtration, and data normalization.

4.4 Algorithm selection for anomaly detection models

In order to determine the model that works the best for the provided PMU data, a variety of models were developed and assessed based on their anomaly detection performances. Within the context of this study, four distinct models, namely: 1. LSTM, 2. Bi-LSTM, 3. CNN and 4. C-LSTM were compared. It should be noted that Bi-LSTM as well as C-LSTM models are hybridized, meaning they consist of two consecutive LSTM and CNN with LSTM layers respectively.

LSTM-based models excel at capturing long-term temporal dependencies within the applied dataset. Thus, may lead to increased accuracies within the domain of PMU-based anomaly detection. One additional benefit of LSTM-based models can be denoted as their robustness towards the vanishing gradient problem (Noh 2021), further strengthening their temporal pattern recognition capabilities. However, it should be noted that LSTM-based models may struggle with spatial dependencies, which is a major limiting factor within the PMU datasets. Additionally, training LSTM-based models may require longer computational times and hence, increased computational requirements.

On the other hand of the spectrum, CNN-based models excel at capturing spatial patterns in a given dataset. As the PMU datasets are both spatiotemporal, CNNs can offer increased benefits within the spatial domain. Additionally, compared to LSTM-based algorithms, CNNs are more efficient, further decreasing the computational requirements. On the downsides of pure CNN-based algorithms, and their limited ability to capture temporal dependencies must be highlighted. Additionally, their increased sensitivity to variations in data shape and size, further requires careful hyper-parameter tuning.

Hybridization offers increased benefits by aiming to combine the strong sides of each algorithm. Thus, mitigating their individual downsides. Within the performed study, two hybrid models, C-LSTM and Bi-LSTM were also assessed. The C-LSTM model combines the individual strengths of both CNN and LSTM models. Thus, can capture spatiotemporal dependencies within the PMU datasets. However, the hybridization can lead to increased

computational capabilities during the training phase. Additionally, careful hyperparameter optimization is similar to pure CNN models. On the other hand, Bi-LSTM models can capture bi-directional temporal dependencies within the training data and thus, lead to a more comprehensive understanding of the sequential data. However, similar to C-LSTM, may need increased computational power during the training phase and may require a larger training dataset.

In summary, LSTMs excel at capturing temporal dependencies within the given datasets and CNNs excel at capturing spatial patterns. By amalgamating both models, a C-LSTM network may excel in both, whereas Bi-LSTM networks may capture bi-directional temporal dependencies better (Sharadga et al. 2020). It should be noted that within the performed study, a single comprehensive PMU dataset was utilized. Hence, the applied models were not trained with additional datasets. Thus, it was not possible to assess their individual performances on various PMU datasets.

4.5 AI-based adversarial noise attacks

The demonstrated experimental framework was designed to evaluate the robustness of four distinct deep-learning models. We subjected these models to several adversarial attack methods to rigorously assess their robustness. These included the BIM, Madry, MIM, and PGD. Each approach challenges the models in a unique way, simulating real-world adversarial scenarios. BIM introduces small iterative perturbations to input data, Madry is an adversarial training-based method, MIM employs momentum for faster perturbation, and PGD computes the worst-case perturbation.

Performed experiments tested different EPS values, determining the level of adversarial perturbations. The range of epsilon values varied from 0.0 to 4.1, encompassing a wide range of adversarial intensities. We analysed how the models reacted to adversarial attacks, ranging from small changes to more significant distortions.

4.6 AI-based cyberattack mitigation

Securing power systems against cyber threats is critical in safeguarding vital infrastructure, particularly power grids. Before delving into the transformative role of AI in augmenting cyber defenses, it is paramount to understand the foundational conventional cybersecurity measures integral to fortifying power systems against cyber threats. Several crucial approaches play a pivotal role in enhancing the resilience of power systems.

Firewalls and intrusion detection systems (IDS) constitute the initial line of defense against cyber threats. Firewalls diligently monitor and regulate incoming and outgoing network traffic based on pre-established security rules. Complementing this, IDS actively surveil network or system activities, swiftly identifying malicious activities or policy violations. Detection of anomalies in network traffic triggers alarms, prompting further investigation and response.

Encryption and implementing secure communication protocols, such as Secure Sockets Layer and Transport Layer Security (SSL/TLS), are instrumental in securing data during transmission. This is essential to guarantee the confidentiality and integrity of information, shielding it from potential eavesdropping and man-in-the-middle attacks. Even if intercepted, the encrypted data remains indecipherable without the requisite decryption keys.

Access control and authentication mechanisms are robust barriers against unauthorized access to critical systems. This involves meticulously defining user roles and permissions, ensuring only authorized personnel can access specific resources. Incorporating multi-factor authentication (MFA) adds a security layer by mandating users to verify their identity through multiple authentication methods using passwords, tokens, or biometrics.

Regular software updates and patch management are imperative to address vulnerabilities cyber adversaries exploit. They consistently keep software and systems up-to-date, aiding in closing security loopholes and maintaining the system's resilience against evolving cyber threats. This applies to operating systems and to, network devices, and any software components integral to power system operations.

Preparedness for cyber incidents is as crucial as preventive measures. Establishing an incident response plan enables organizations to detect, respond to, and recover from cyber incidents effectively. Concurrently, regular cybersecurity training programs for personnel enhance their awareness of potential threats, equipping them with the knowledge to identify and respond to security incidents promptly.

Network segmentation emerges as a strategic defense mechanism by dividing a network into segments or zones. This containment strategy adds a defense layer, limiting potential threats within isolated areas. Even if an attacker gains access to one segment, network segmentation prevents lateral movement, denying the impact of the intrusion.

Regularly backing up critical data and implementing robust disaster recovery plans include vital measures to ensure organizations can swiftly restore operations during a cyberattack. This proactive approach mitigates potential damage caused by data loss or system disruptions, reinforcing the resilience of power systems against cyber threats.

4.7 Algorithm training and testing

This section further specifies the steps taken during the training, optimization, and testing steps of the various AI models utilized within the context of the performed research.

4.7.1 Algorithm 1: LSTM training

The LSTM model was constructed through the following algorithm. As the first step, the architecture of the LSTM network was established, comprising two LSTM layers, each containing 64 units. This was followed by a dense layer equipped with a sigmoid activation function for binary classification. Secondly, crucial parameters such as the learning rate, batch size, and the desired number of training epochs were defined to facilitate the training process. Subsequently, the LSTM model's parameters were initialized randomly to commence the training process. As the fourth step within the algorithm development, the training data was then iterated over for the specified number of epochs, where each iteration involved the forward pass of the LSTM model to predict outcomes, and the calculation of the loss through binary cross-entropy computation.

Next, the backpropagation was executed upon loss calculation, which facilitated the adjustment of model parameters using the Adam optimizer in order to optimize the model's predictive capacity. Thus, this iterative training process which encompasses steps four to six was repeated until the convergence was achieved, or the specified number of training epochs was reached. As a summary, the steps performed during the model training can be given as:

1. Initialize the LSTM network architecture with two LSTM layers, each consisting of 64 units, followed by a dense layer with sigmoid activation for binary classification.
2. Set the learning rate, batch size, and number of epochs for training.
3. Initialize the LSTM model parameters randomly.
4. Iterate through the training data for the specified number of epochs.
5. Calculate the forward pass of the LSTM model and compute the loss using binary cross-entropy.
6. Perform backpropagation to update the model parameters using the Adam optimizer.
7. Repeat steps 4–6 until convergence or the specified number of epochs is reached.
8. Evaluate the trained LSTM model on the testing set using the selected performance metrics.

Finally, the trained LSTM model was assessed using the testing portion of the dataset, where selected performance metrics were computed and evaluated.

As can be seen, steps four to six correspond to hyperparameter optimization. The individual results of the parameters after the tuning process are illustrated within Table 2.

4.7.2 Algorithm 2: CNN training

For the training process of the utilized CNN model, the following steps were performed: Firstly, the CNN architecture was established which featured two convolutional layers where each layer was composed of 32 filters with dimensions of 3×3. This was succeeded by max-pooling layers and a final dense layer which was intended for classification purposes. Subsequently, key parameters such as the learning rate, batch size, and the desired number of training epochs were defined to facilitate the model's training process.

Following this, the CNN model's parameters were initialized randomly as a starting point for the training step. The training data was then iterated over for the specified number of epochs, encompassing forward passes of the CNN model to generate predictions and computations of loss via binary cross-entropy. With the loss computed, backpropagation was performed in order to enable the adjustment of the model's parameters by utilizing the Adam optimizer, Therefore, enhancing the model's predictive capabilities.

This iterative training process, which involved the steps four through six was then repeated until either the model was converged or the predefined number of training epochs was met. The results after this hyperparameter tuning can be found within Table 3. For ease of reproducibility, the summary of the training process can be denoted as:

1. Define the CNN architecture with two convolutional layers, each consisting of 32 filters of size 3×3, followed by max-pooling layers and a dense layer for classification.
2. Specify the learning rate, batch size, and number of epochs for training.
3. Initialize the CNN model parameters randomly.

Table 2 Results of hyperparameter tuning for the assessed LSTM algorithm

LSTM Model Parameters		
Parameter	Type	Optimized Value
LSTM	Hidden units	32
Dropout	Dropout value	0.12
Dense	Units	1

Table 3 Results of hyperparameter tuning for the assessed CNN algorithm

CNN model parameters		
Parameter	Type	Optimized value
Conv1D	Filters	32
	Kernel Size	3
	Activation function	relu
Conv1D	Filters	16
	Kernel Size	3
	Activation function	relu
Conv1D	Filters	64
	Kernel Size	3
	Activation function	relu
Max Pooling	Pool Size	2
Dense	Units	50
	Activation	relu
Dense	Units	1

4. Iterate through the training data for the specified number of epochs.
5. Perform a forward pass of the CNN model and compute the loss using binary cross-entropy.
6. Utilize backpropagation to update the model parameters using the Adam optimizer.
7. Repeat steps 4–6 until convergence or the specified number of epochs is reached.
8. Evaluate the trained CNN model on the testing set using the selected performance metrics.

Finally, the trained CNN model was evaluated using the testing dataset, where its performance was assessed based on chosen performance metrics.

4.7.3 Algorithm 3: C-LSTM training

For the training purposes of the C-LSTM model, the subsequent procedure was employed: To begin with, the architecture of the model was defined, which encompassed the formulation of a CNN with two layers, each comprising 32 filters sized at 3×3. These were succeeded by max-pooling layers to extract key features from the data. Following the definition of the CNN architecture, the LSTM network was constructed which incorporated two LSTM layers, each containing 64 units. Later on, this was followed by a dense layer that utilized a sigmoid activation function for binary classification purposes.

As the next step, the essential hyperparameters such as the learning rate, batch size, and the desired number of training epochs were specified. Later on, both the CNN and LSTM models' parameters were initialized randomly. The training portion of the dataset was then iterated through for the predetermined number of epochs. For each iteration, the forward pass of the CNN model was executed, and the loss was computed using binary cross-entropy. Subsequently, the forward pass of the LSTM model was performed and its loss was calculated in a similar manner.

In order to optimize both of the models which are working in tandem, backpropagation was employed, adjusting the parameters for both models through the use of the Adam optimizer. This iterative process, which encompasses the performed steps through four and

Table 4 Results of hyperparameter tuning for the assessed C-LSTM algorithm

C-LSTM model parameters		
Parameter	Type	Optimized value
Conv1D	Filters	64
	Kernel Size	3
	Activation	relu
LSTM	Hidden units	64
Dropout	Dropout value	0.12
LSTM	Hidden units	32
Dropout	Dropout value	0.12
Dense	Units	1

eight was then repeated until either the convergence was achieved, or the specified number of training epochs were reached. Upon completion, the hybrid model was then evaluated using the testing portion of the dataset by employing the chosen performance metrics in order to see the resulting accuracies. As a summary and a guide for reproducibility, the mentioned steps can further be denoted as the following. Additionally, the results of hyperparameter optimization can be found in Table 4.

1. Define the CNN architecture with two layers, each consisting of 32 filters of size 3x3, followed by a max-pooling layer.
2. Initialize the LSTM network architecture with two LSTM layers, each consisting of 64 units, followed by a dense layer with sigmoid activation for binary classification.
3. Specify the learning rate, batch size, and number of epochs for training.
4. Initialize both CNN and LSTM model parameters randomly.
5. Iterate through the training data for the specified number of epochs.
6. Perform a forward pass of the CNN model and compute the loss using binary cross-entropy.
7. Calculate the forward pass of the LSTM model and compute the loss using binary cross-entropy.
8. Utilize backpropagation to update the model parameters using the Adam optimizer.
9. Repeat steps 4–6 until convergence or the specified number of epochs is reached.
10. Evaluate the trained hybridized model on the testing set using the selected performance metrics.

4.7.4 Algorithm 4: Bi-LSTM training

For the training procedure of the Bi-LSTM model, the following steps were employed within the context of this research: As the initial step, two LSTM network architectures were initialized where each architecture was comprised of two LSTM layers, each containing 64 units. In the second network, an additional dense layer was added in order to make the resulting hybrid model suitable for classification purposes. Following the architecture initialization, various essential parameters such as the learning rate, batch size, and the desired number of training epochs were established in order to facilitate the training

Table 5 Results of hyperparameter tuning for the assessed Bi-LSTM algorithm

Bi-LSTM model parameters		
Parameter	Type	Optimized value
LSTM	Hidden units	32
Dropout	Dropout value	0.12
Dense	Units	1

process. Subsequently, the LSTM model's parameters were randomly initialized in order to create a starting point for training purposes.

The training portion of the data was then iterated over for the specified number of epochs, during which the forward pass of the LSTM model was computed together with the resulting loss via binary cross-entropy. Upon the loss calculation, backpropagation was employed in order to update the model parameters by using the Adam optimizer. This iterative process was then repeated until the convergence was achieved or the initially specified number of epochs was reached. The results of the hyperparameter tuning after this iterative process are denoted in Table 5. Meanwhile, the step-by-step guide for the overall training process can further be given as:

1. Initialize 2 LSTM network architecture with two LSTM layers, each consisting of 64 units, followed by a dense layer in 2nd network for classification purposes.
2. Set the learning rate, batch size, and number of epochs for training.
3. Initialize the LSTM model parameters randomly.
4. Iterate through the training data for the specified number of epochs.
5. Calculate the forward pass of the LSTM model and compute the loss using binary cross-entropy.
6. Perform backpropagation to update the model parameters using the Adam optimizer.
7. Repeat steps 4–6 until convergence or the specified number of epochs is reached.
8. Evaluate the trained LSTM model on the testing set using the selected performance metrics.

Finally, the trained Bi-LSTM model was then evaluated on the testing portion of the dataset, where the resulting performance of the hybridized model was assessed based on the selected performance metrics.

4.8 Evaluation and performance metrics

For the AI attacks section, there are numerous approaches for evaluating the suggested methods and their predictions. The effectiveness of the evaluation procedure is measured in terms of its capacity to identify presented anomalies when used for anomaly detection. Given that the LSTM approach predicts future values, it only makes sense to compare the anticipated values to the actual measurements taken by the PMUs using an error vector.

In order to evaluate the proposed algorithms, a dynamic error thresholding technique was used with a standard x-sigma threshold. This technique was first developed by NASA in Pukelsheim (1994). However, the approached used in this research will base itself into a modified version as performed in Tinawi (2019).

The first step for evaluation is to find the prediction error, which is done by looking at individual datapoints directly. The prediction error can be computed as:

$$e = |y_t - \hat{y}_t| \tag{11}$$

However, the modified approach as done in Tinawi (2019) utilizes a series of errors in order to determine the threshold, which is done by defining a window size and a number of errors to combine in a given error vector $\mathbf{e} = [e_{t-n}, \ldots, e_{t-1}, e_t]$, where n denotes the window size.

The LSTM-based methods tend to experience error spikes frequently (Hundman et al. 2018). Thus, Exponentially Weighted Moving Average (EWMA) was used in order to smoothen out the error vector with the formula below where alpha denotes how quickly the weights tend to zero, while propagating back in time.

$$EWMA_t = \alpha * e_t + (1 - \alpha) * EWMA_{t-1} \tag{12}$$

The smoothed error vector, denoted as $\mathbf{e}_s$, is then to be utilized for calculating the error threshold. The threshold, denoted as ϵ is then selected to mitigate the amount of anomalies that is marked, reducing the false positives. However, there needs to be a balance between the false negative and false positive values. Therefore, the threshold vector can be modeled as:

$$\epsilon = \mu(\mathbf{e}_s) + \mathbf{z}\sigma(\mathbf{e}_s) \tag{13}$$

where μ and σ denote the mean and the standard deviation of the smoothed error vector respectively, whereas $\mathbf{z}$ is the positive number that is chosen in order to scale up the threshold. Therefore, a lower value of $\mathbf{z}$ will result in a lower anomaly detection threshold, resulting in more false positives. As stated in Hundman et al. (2018), a number between 2 and 10 was used for this research.

Furthermore, ϵ was chosen as and modeled as:

$$\epsilon = argmax(\epsilon) = \frac{\Delta\mu(\mathbf{e}_s)/\mu(\mathbf{e}_s) + \Delta\sigma(\mathbf{e}_s)/\sigma(\mathbf{e}_s)}{|\mathbf{e}_a| + |\mathbf{E}_{seq}|^2} \tag{14}$$

where

$$\begin{aligned} \Delta\mu(\mathbf{e}_s) &= \mu(\mathbf{e}_s) - \mu(\{e_s \in \mathbf{e}_s | e_s < \epsilon\}) \\ \Delta\sigma(\mathbf{e}_s) &= \sigma(\mathbf{e}_s) - \sigma(\{e_s \in \mathbf{e}_s | e_s < \epsilon\}) \\ \mathbf{e}_a &= \{e_s \in \mathbf{e}_s | e_s > \epsilon\} \end{aligned}$$

and $\mathbf{E}_{seq}$ denotes the continuous sequence of $e_a \in \mathbf{e}_a$. The severity of the detected anomaly is also an important factor. Therefore, an anomaly score is assigned to each detected anomaly via Eq. 15:

$$s^{(i)} = \frac{max(\mathbf{e}_{seq}^{(i)} - argmax(\epsilon)}{\mu(\mathbf{e}_s) + \sigma(\mathbf{e}_s)} \tag{15}$$

Meanwhile, for the algorithms (LSTM, CNN, Bi-LSTM and C-LSTM) that are responsible for anomaly detection within the PMU domain, three evaluation metrics were utilized. For these algorithms, a Mean Square Error (MSE) was used as a loss function during the training step due to the fact that it heavily punishes outliers. Thus, providing better results for detecting anomalies. The equation for MSE can be denoted as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{16}$$

In the Eq. 16, the n denotes the number of data points that are predicted, whereas x_i denotes the real measured values. Meanwhile, the notation $\hat{x}_i$ represents the predicted values by the assessed algorithms. In order to assess the performance of the models, precision, recall and F1 scores were utilized as performance metrics. The precision score represents the proportion of true positives with the total points identified as positive. Thus, anomalies that are identified by a model with a high precision score are more likely to be actual anomalies. Meanwhile, the recall score represents the proportion of true positives divided by the sum of true positives and false negatives. Thus, a model with a high recall score is more adept at finding positive cases, albeit at the potential cost of misclassifying some negative instances as positive. Finally, the F1 score combines both precision and recall, as it represents the harmonic mean of both mentioned scores. The equations for the utilized performance metrics can be denoted as:

$$Precision = \frac{N_{TruePositives}}{N_{TruePositives} + N_{FalsePositives}} \tag{17}$$

$$Recall = \frac{N_{TruePositives}}{N_{TruePositives} + N_{FalseNegatives}} \tag{18}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{19}$$

4.9 Validation

To validate the effectiveness of the proposed methodology, the results obtained from the LSTM and CNN models will be compared with existing anomaly detection approaches in power systems. Additionally, sensitivity analyses will be conducted to assess the robustness of the models to variations in input parameters and to evaluate their performance under different operating conditions.

5 Experimental results

This section is further subdivided into 2 categories. In the first subsection, the results for the developed anomaly detection models will be highlighted and discussed. Meanwhile, in the second subsection, the results of the adversarial attacks will be highlighted and discussed.

5.1 Anomaly detection results

The outcomes of the assessed anomaly detection algorithms are detailed in Table 6. As can be seen from the performance metrics results, the models diverge significantly in terms of

Table 6 Performance results for the different models

Model	Noise filtration	Recall (%)	Precision (%)	F1-Score(%)
CNN	Yes	97.50	86.70	90.70
	No	81.25	94.20	89.65
LSTM	Yes	93.75	97.40	95.54
	No	98.41	77.50	86.71
Bi-LSTM	Yes	98.75	94.04	96.34
	No	85.0	90.66	87.74
C-LSTM	Yes	95.50	98.70	96.81
	No	94.76	95.14	96.84

anomaly detection capacity across the metrics. The CNN model with noise filtration exhibited a behavior to classify an excessive number of points as anomalies, which impacts both the precision and F1 score. However, this tendency also leads to a scarcity of false negatives, with a recall score of 94.20%. Meanwhile, the CNN model trained with the non-noise filtered data identified a limited number of anomalies, which was deemed to be the result of heightened noise levels during the training section, which in return led to a higher threshold for anomaly detection. From the Fig. 9, it can be seen that the pure CNN struggles with false positives, especially after the FDI attack as the algorithm needs time to settle back to normal anomaly threshold levels after the attack has been performed.

Within the LSTM-based models, the utilization of noise filtration within the training dataset proves indispensable in order to achieve better accuracy scores for anomaly detection. As can be seen from the Table 6, the LSTM model surpasses the CNN model when the noise filtration is applied during the training phase. However, exhibits low performance compared to CNN when the noise filtration is omitted. With the noise filtration of training data, the F1-score of LSTM algorithm improves to 92.16 from 78.79%. One additional important finding can be found in Figs. 10 and 12. In both figures, it is evident that both

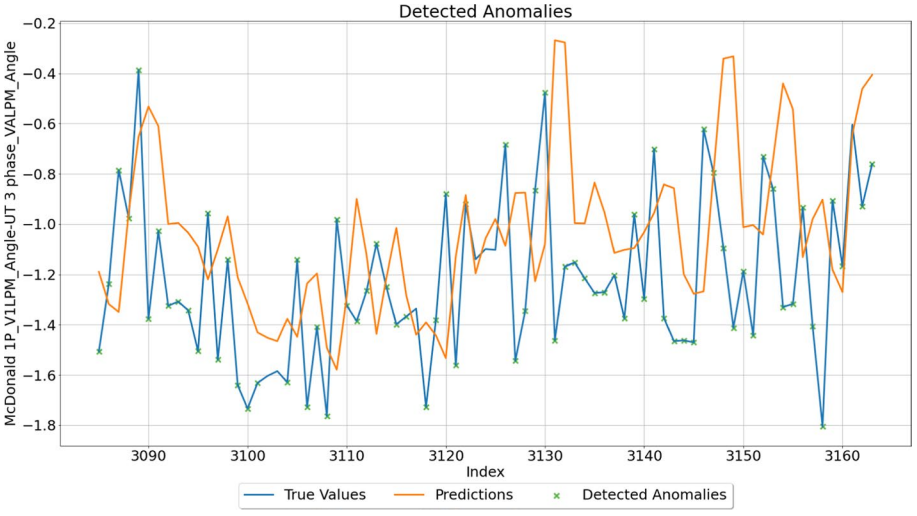

Fig. 9 The CNN model without noise filtration

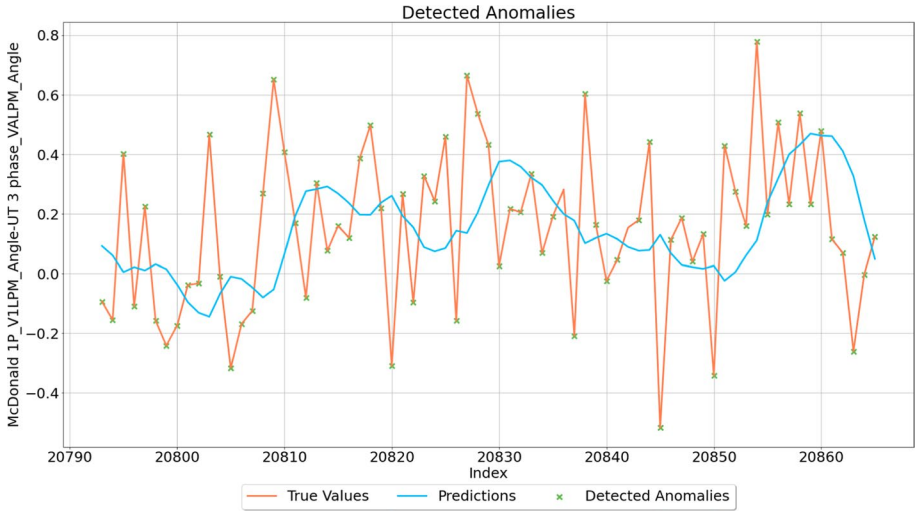

Fig. 10 LSTM—detected anomalies with pred with noise filter

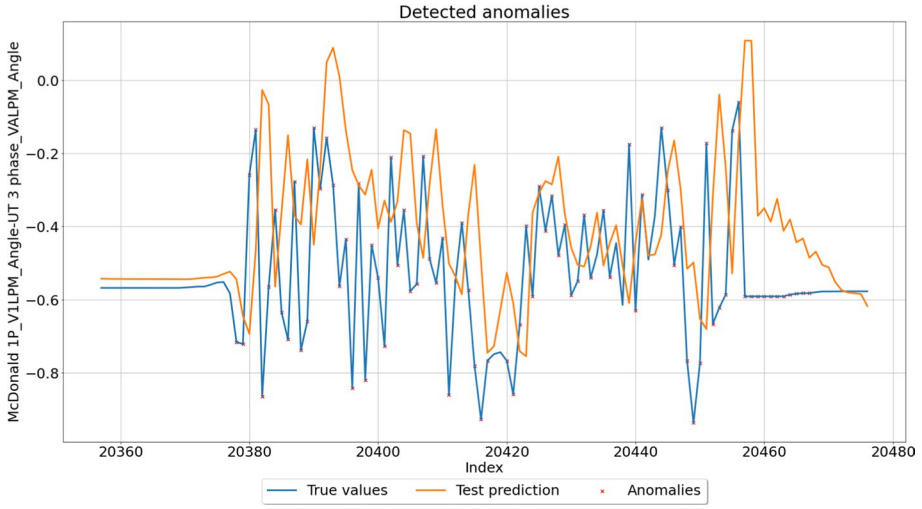

Fig. 11 C-LSTM—detected anomalies with noise filtration

pure and BiLSTM have a bad fit compared to the ground truth data. However, both of their performance exceeds an F1-score of 90%.

Meanwhile, the hybridization of the LSTM with the CNN yields the best results by combining the advantages of both algorithms. From the Fig. 11, the results of the hybrid cLSTM can be observed. It is evident that despite the overshoots in the model predictions compared to the ground truth data, this hybrid model had the best results among the assessed models, both with and without noise filtration during the training phase. The F1-score of the hybrid cLSTM algorithm can be seen in Table 6 as 96.89%.

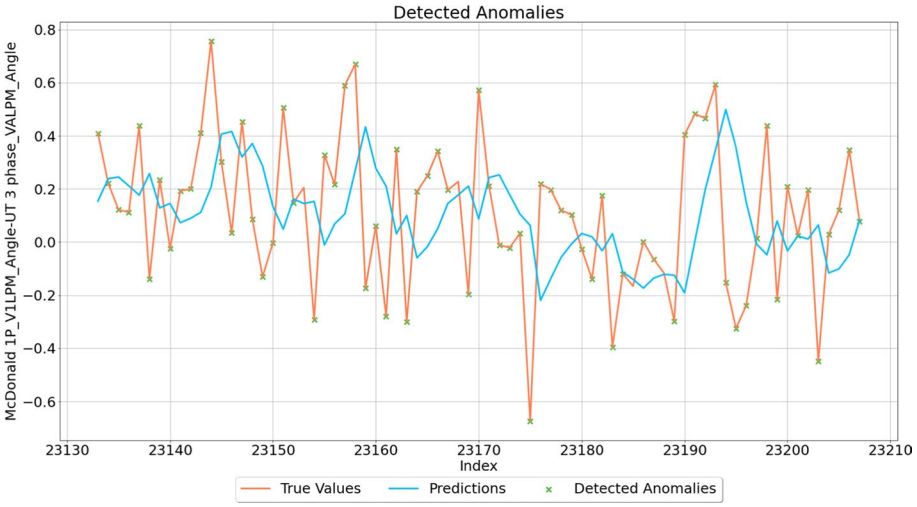

Fig. 12 Bi-LSTM—detected anomalies with noise filtration

5.2 Adversarial attack results

In this section, we will discuss the results of the experiments conducted to evaluate the robustness of different AI models (LSTM, cLSTM, biLSTM, and CNN) in the context of anomaly detection for power systems when subjected to adversarial attacks with varying EPS values. The metrics used for evaluation include prediction performance metrics, focusing on the "Undefended" and "Defended" models, which represent the models' performance without and with defenses against adversarial attacks, respectively. For the LSTM model, the results show a clear trend as the EPS value increases. In the "Undefended" model, the prediction performance metrics start with relatively high values at an EPS of 0.41 but steadily deteriorate as EPS increases. The "Defended" model, which incorporates countermeasures against adversarial attacks, exhibits a significant improvement in prediction performance compared to the "Undefended" model. The defended model consistently outperforms the undefended one across all EPS values, with a substantial reduction in prediction errors. In the case of the cLSTM model, we observe a similar pattern in the "Undefended" model's performance as the EPS value increases. Prediction performance metrics start at reasonable levels at an EPS of 0.41 but degrade with higher EPS values. The "Defended" model for cLSTM also exhibits a clear improvement over the "Undefended" model, indicating the effectiveness of the defenses against adversarial attacks. The defended model consistently outperforms the undefended one, with a noticeable reduction in prediction errors. The results for the biLSTM model indicate a more varied performance as the EPS value changes. In the "Undefended" model, prediction performance metrics show fluctuations as EPS values increase, with some values improving while others worsen. In contrast, the "Defended" model consistently improves prediction performance across different EPS values. It is noteworthy that for specific EPS values, the defended model outperforms the undefended one by a substantial margin. For the CNN model, the results show that the "Undefended" model's performance deteriorates with increasing EPS values, as evidenced by a rise in prediction errors. The "Defended" model, on the other hand, consistently performs better than the "Undefended" model across all

EPS values, with a substantial reduction in prediction errors. The improvement in prediction performance is particularly pronounced at higher EPS values.

Our study utilizes MSE as the primary evaluation metric for assessing the performance of anomaly detection models. MSE is particularly appropriate for our task as it is susceptible to outliers, which aligns well with detecting anomalies characterized by significant deviations from normal patterns. This metric ensures that our models effectively capture and quantify deviations in PMU measurements, providing a transparent and interpretable measure of prediction accuracy for continuous data.

The experiments highlight the vulnerability of AI models in power system anomaly detection to adversarial attacks, with prediction performance degrading as EPS values increase. However, the introduction of defenses significantly enhances the models' resilience, leading to improved prediction accuracy and reliability, especially at higher EPS values. These findings underscore the importance of developing and implementing robust countermeasures to protect power systems from adversarial threats and maintain the security and reliability of critical infrastructure. The central objective of our experimental inquiry was to evaluate the robustness of four distinct computational models: Long Short-Term Memory (LSTM), convolutional LSTM (cLSTM), bidirectional LSTM (biLSTM), and Convolutional Neural Network (CNN). This assessment was conducted in the context of a diverse array of adversarial attack methodologies. Our primary aim was to scrutinize the models' performance under a spectrum of adversarial perturbations parameterized by epsilon (EPS) values. The performance evaluation metric of choice was the Mean Squared Error (MSE), which quantifies the dissonance between the predictive outcomes of the models and the verifiable ground truth labels. The outcomes of our extensive experimentation are thoughtfully encapsulated within Tables 8, 9, 10, and 11, with an additional visual representation offered by Fig. 13. Each table contains empirical insights, displaying MSE values from various adversarial strategies. These stratagems encompass the Basic Iterative Method (BIM), Madry, Momentum Iterative Method (MIM), and Projected Gradient Descent (PGD) and are executed across a gamut of epsilon values.

The specific parameter settings for these attacks are summarized in Table 7.

The robustness of each model was evaluated by comparing the MSE before and after the application of adversarial attacks. Lower increases in MSE indicate better robustness.

To provide an accurate reference, we use MSE (Normal) to symbolize the baseline MSE calculated based on unaltered data under clean conditions. The columns on either side of this reference in the tables represent the MSE values obtained after the models were exposed to various adversarial attack methods. If one looks at Table 8, we carefully examine the performance of the LSTM model. The rows in this table indicate the epsilon values that show the level of adversarial perturbation introduced. The MSE(Normal) column represents the model's baseline performance when dealing with uncontaminated data. The other columns show the MSE values when the model is exposed to different adversarial strategies. Tables 9, 10, and 11 show the MSE values for the cLSTM, biLSTM, and CNN

Table 7 Parameter settings for adversarial attacks

Attack method	Iteration count	Step size	Epsilon (ϵ)
BIM	10	0.01	0.1
Madry Attack	40	0.01	0.1–0.3
MIM	10	0.01	0.1
PGD	40	0.01	0.1–0.3

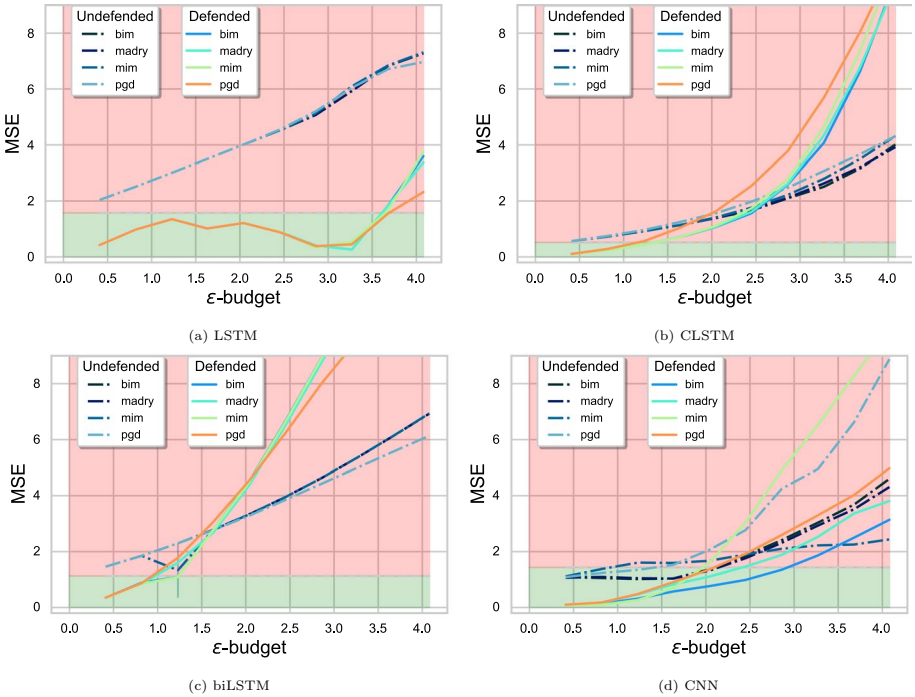

Fig. 13 MSE values for assessed attack types across the utilized algorithms for defended and undefended models

models across different epsilon variations and adversarial attacks. The MSE is a useful tool that helps us analyze and understand the results of our experiments. It shows how well our models can withstand challenges and unexpected events. A lower MSE value means that our models are more resilient and can handle disruptions without deviating significantly from our expected predictions. On the other hand, a higher MSE value indicates that our models are more vulnerable and may not accurately predict outcomes when faced with unexpected events.

Similar patterns can be observed in Tables 9, 10, and 11. Each table provides detailed information about the results of the cLSTM, biLSTM, and CNN models, respectively. The tables carefully record the MSE values for different epsilon values and types of attacks, giving insight into how these models respond to adversarial attacks. Through these experiments, we have discovered a wealth of empirical findings that provide valuable insights into how models can better withstand attacks. The MSE judges the model's performance, with a lower value indicating a higher level of resilience against adversarial perturbations. This demonstrates the model's ability to persevere in challenging situations and stay on track towards expected outcomes. Conversely, an elevated MSE value reveals vulnerabilities in the model's predictive capabilities and highlights the impact of adversarial forces. We can gain valuable insights by carefully analyzing the MSE values across a range of attack scenarios. This allows us to identify each model's unique strengths and weaknesses and helps us develop effective defense mechanisms. Armed with this knowledge, we can work to fortify our models against adversarial attacks. These findings shed light on the path towards creating machine learning models that are strong and genuinely resilient.

The robustness of four distinct models—LSTM, cLSTM, biLSTM, and CNN—underwent experimentation to assess their performance against various adversarial attack methods. These experiments yielded insightful findings regarding the models' ability to withstand different levels of adversarial perturbations. Mean Squared Error (MSE) values were calculated for each model and attack method combination, revealing their strengths and weaknesses concerning adversarial examples. Models with lower MSE values exhibited greater resilience, while those with higher values were deemed more vulnerable. These findings can be utilized to develop effective defense mechanisms that enhance the models' robustness and reliability in practical applications. By comprehending the impact of different attack methods and corresponding model responses, researchers can devise strategies to improve the models' ability to withstand adversarial attacks, ensuring their reliability and trustworthiness. The knowledge gained from these experiments will contribute to the advancement of adversarial robustness research in the field of machine learning and assist in the development of more secure and dependable models.

6 Conclusion and discussions

The experimental results demonstrate the vulnerability of the LSTM, cLSTM, biLSTM, and CNN models to adversarial attacks. The MSE values obtained under various attack methods and epsilon values highlight the models' susceptibility to adversarial perturbations. Higher MSE values indicate increased deviation from the expected outputs and signify a more significant impact of the attacks on the models' predictions. The findings suggest that the trustworthiness of the models varies across different attack methods and epsilon values. The Madry attack shows promising results in generating strong adversarial examples that generalize well across models and defenses. On the other hand, the BIM, MIM, and PGD attacks also pose significant threats to the models' performance, exhibiting increased MSE values as the magnitude of the adversarial perturbations (epsilon) increases. Our findings indicate that LSTM models consistently achieve the lowest MSE values against all four adversarial attacks, demonstrating their superior ability to generalize in the presence of perturbations. This is likely due to LSTM models' ability to capture long-range dependencies in the input data, which helps them to better understand the underlying relationships between inputs and outputs. BiLSTM models generally have higher MSE values than LSTM and cLSTM models against BIM and MIM attacks, but they exhibit comparable performance against Madry and PGD attacks. This suggests that biLSTM models are more effective at handling additive perturbations compared to multiplicative perturbations. These results emphasize the importance of developing effective defense mechanisms to enhance the models' robustness against adversarial attacks. Future research efforts should focus on exploring and implementing advanced defense techniques, such as adversarial training, ensemble methods, and model distillation. These techniques aim to improve the models' ability to withstand adversarial perturbations and maintain accurate predictions even in the presence of sophisticated attacks. Additionally, investigating the transferability of adversarial examples across different models and datasets would provide valuable insights into the generalization capabilities of the attacks. Furthermore, exploring the impact of different hyperparameters, such as learning rates, batch sizes, and network architectures, on the models' vulnerability to adversarial attacks could yield further insights into improving robustness. One area for improvement of this study is its reliance on historical data,

which may not capture the full range of possible anomalies that could occur in real-time operation. Additionally, the performance of the anomaly detection algorithms may vary depending on the specific characteristics of the power grid and the types of anomalies present. It should be noted that the generalizability of the results to other power systems should be considered with caution. The study's results show that looking into the balance between how accurate AI-based anomaly detection models are and how safe they are against possible cyberphysical attacks for the critical infrastructure being thought about is essential. This article also shows how different AI models can be used together and in order, both at the same time and against each other, for many different tasks, such as finding strange things, carrying out cyberattacks, and preventing them. Overall, this study sheds light on the importance of understanding and mitigating the vulnerabilities of AI models to adversarial attacks. By developing more robust and resilient models, we can enhance the reliability and trustworthiness of AI systems in real-world applications.

7 Future work

In future research, several avenues can be explored to advance the understanding and defense against adversarial attacks. Here are some potential directions:

Adversarial training with diverse attack scenarios: Extending the current experiments by incorporating a more comprehensive range of adversarial attacks, such as Carlini and Wagner attack or Jacobian-based Saliency Map Attack, can provide a more thorough evaluation of the model's robustness. Adversarial training on these diverse attack scenarios can lead to the developing of more resilient models. *Defense mechanisms:* Investigating and developing novel defense mechanisms, including gradient regularization, input preprocessing techniques, and randomized smoothing, can contribute to creating models more resistant to adversarial perturbations. Evaluating the effectiveness of these defenses against a broader range of attack methods is crucial. *Real-world applications:* Extending the evaluation of adversarial robustness to real-world applications, such as autonomous driving or medical diagnosis, can help assess the models' performance under more practical and complex scenarios. This research can lead to the development of AI systems that are reliable and secure in critical domains. *Adversarial detection and explainability:* Designing techniques for detecting adversarial examples and explaining the model's decision-making process in the presence of such examples are essential for building trust in AI systems. Exploring methods to identify and interpret adversarial attacks can help develop more transparent and accountable AI models. By pursuing these avenues of research, we can enhance the robustness of machine learning models against adversarial attacks and foster the development of reliable and secure AI systems in various domains.

Integrating emerging technologies such as quantum and edge computing could significantly improve processing capabilities and reduce the latency of anomaly detection systems. Expanding the application of these AI models to other critical infrastructure sectors, such as transportation and healthcare, could improve resilience against anomalies and cyber-attacks. Implementing adaptive learning mechanisms to allow models to evolve with new anomalies and threats could enhance long-term effectiveness. Investigating the ethical implications and security vulnerabilities of deploying AI-based anomaly detection in critical infrastructures is also essential. Ensuring these systems are

effective, secure, and compliant with privacy regulations is crucial. Addressing these research questions can advance anomaly detection in critical infrastructure systems, enhancing their resilience and security against adversarial threats.

Appendix

See Table 8, 9, 10, and 11.

Table 8 LSTM

	BIM		MIM		MADRY		PGD	
EPS	**Undef.**	**Defended**	**Undef.**	**Defended**	**Undef.**	**Defended**	**Undef.**	**Defended**
0.41	2.038603	0.424041	2.038603	0.424039	2.038603	0.424032	2.038603	0.424030
0.82	2.510237	0.978672	2.510237	0.978562	2.510237	0.978307	2.510237	0.978600
1.23	3.006264	1.350685	3.006263	1.348693	3.006263	1.345158	3.006263	1.349458
1.63	3.501387	1.017411	3.501387	1.016827	3.501388	1.018399	3.501388	1.016745
2.04	4.011350	1.212153	4.011289	1.212072	4.011016	1.213226	4.011215	1.212370
2.45	4.519428	0.899939	4.519774	0.899127	4.519058	0.900515	4.542234	0.884025
2.86	5.070283	0.402693	5.071837	0.400386	5.130139	0.340077	5.199975	0.386309
3.27	5.927121	0.261244	5.927875	0.260158	6.097188	0.536793	6.042330	0.448423
3.68	6.829479	1.839760	6.826449	1.837342	6.841919	1.665468	6.712505	1.554542
4.08	7.309332	3.597418	7.276105	3.373633	7.309711	3.752743	6.968611	2.320992

Table 9 cLSTM

	BIM		MIM		MADRY		PGD	
EPS	Undef.	Defended	Undef.	Defended	Undef.	Defended	Undef.	Defended
0.41	0.567433	0.103676	0.569134	0.106258	0.567307	0.103674	0.569535	0.106788
0.82	0.727147	0.232114	0.728438	0.235902	0.726714	0.232192	0.750101	0.285447
1.23	0.917919	0.432103	0.918342	0.433747	0.917765	0.431965	0.964079	0.562987
1.63	1.132826	0.698303	1.136438	0.709684	1.132608	0.697960	1.241917	1.030655
2.04	1.386425	1.068960	1.388759	1.078858	1.407622	1.106707	1.556927	1.619686
2.45	1.661776	1.564339	1.688646	1.616520	1.763487	1.737316	1.977576	2.544655
2.86	2.097034	2.586703	2.115149	2.609740	2.218876	2.782368	2.473083	3.777057
3.27	2.508255	4.075356	2.628392	4.337136	2.792792	4.642182	3.071305	5.691321
3.68	3.156483	6.631951	3.207705	6.768409	3.499580	7.366041	3.668923	8.043638
4.08	4.012347	9.904355	3.920264	9.756875	4.321109	10.597324	4.313021	10.626999

Table 10 biLSTM

	BIM		MIM		MADRY		PGD	
EPS	Undef.	Defended	Undef.	Defended	Undef.	Defended	Undef.	Defended
0.41	1.460959	0.350573	1.461585	0.351523	1.460959	0.349393	1.461768	0.351935
0.82	1.852657	0.887757	1.853901	0.851057	1.852648	0.847769	1.855811	0.872957
1.23	1.333940	1.092849	2.296755	1.607440	1.333913	1.089397	2.296225	1.778938
1.63	2.779864	2.777312	2.781956	2.681719	2.779762	2.774046	2.770984	3.061738
2.04	3.335863	4.510283	3.337653	4.384780	3.335725	4.488595	3.289842	4.551433
2.45	3.954246	6.665024	3.956228	6.517244	3.954098	6.641309	3.853755	6.239162
2.86	4.634205	8.932697	4.634704	8.779609	4.634131	8.920742	4.429744	8.019218
3.27	5.369992	11.070434	5.368169	10.927076	5.369952	11.051568	5.006110	9.598551
3.68	6.150044	12.909105	6.146178	12.740335	6.150092	12.897627	5.584798	10.927830
4.08	6.937686	14.391960	6.930556	14.237649	6.937732	14.390517	6.137514	12.151557

Table 11 CNN

	BIM		MIM		MADRY		PGD	
EPS	Undef.	Defended	Undef.	Defended	Undef.	Defended	Undef.	Defended
0.41	1.089005	0.085219	1.066399	0.100179	1.113065	0.083216	1.086545	0.101014
0.82	1.055095	0.136559	1.094858	0.167576	1.378529	0.092103	1.237819	0.180064
1.23	1.015702	0.317642	1.048992	0.469688	1.607382	0.271097	1.347260	0.484505
1.63	1.039043	0.573274	1.035431	0.822781	1.595004	0.692889	1.527915	0.907190
2.04	1.364652	0.765247	1.320619	1.107039	1.672450	1.598169	2.064729	1.362050
2.45	1.853013	0.987222	1.764516	1.465397	1.901406	3.041248	2.770632	1.920396
2.86	2.413224	1.351913	2.315255	1.883981	2.102646	4.905322	4.247408	2.596251
3.27	3.036348	1.864545	2.931976	2.530704	2.223555	6.529574	4.949097	3.298850
3.68	3.683633	2.500685	3.517498	3.367022	2.254240	8.279675	6.622675	4.027385
4.08	4.603741	3.139967	4.310988	3.811644	2.437570	9.928660	8.877693	4.977233

Author contributions U.C., F.O.C. and U.H. wrote the main manuscript text. U.C., and F.O.C. supervised. U.C. conceptual development. F.O.C. experimental design. U.H. editing

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital)

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abur A, Exposito AG (2004) Power system state estimation: theory and implementation. CRC Press, Boca Raton

Ahmed M, Mahmood AN, Hu J (2016) A survey of network anomaly detection techniques. J Netw Comput Appl 60:19–31

Amini A, Kanfoud J, Gan T-H (2022) An artificial intelligence neural network predictive model for anomaly detection and monitoring of wind turbines using scada data. Appl Artif Intell. https://doi.org/10.1080/08839514.2022.2034718

Amutha A, Uthra RA, Roselyn JP, Brunet RG (2021) Anomaly detection in multivariate streaming pmu data using density estimation technique in wide area monitoring system. Expert Syst Appl 175

Arefin AA, Baba M, Singh NSS, Nor NBM, Sheikh MA, Kannan R, Abro GEM, Mathur N (2022) Review of the techniques of the data analytics and islanding detection of distribution systems using phasor measurement unit data. Electronics 11(18):2967

Ashrafuzzaman M, Chakhchoukh Y, Jillepalli AA, Tosic PT, de Leon DC, Sheldon FT, Johnson BK (2018) Detecting stealthy false data injection attacks in power grids using deep learning, In: 2018 14th international wireless communications & mobile computing conference (IWCMC), IEEE, pp 219–225

Badrinath Krishna V, Weaver GA, Sanders WH (2015) Pca-based method for detecting integrity attacks on advanced metering infrastructure, In: Quantitative evaluation of systems: 12th international conference, QEST 2015, Madrid, Spain, September 1–3, proceedings 12, Springer, 2015, pp 70–85

Badrinath Krishna V, Iyer RK, Sanders WH (2016) Arima-based modeling and validation of consumption readings in power grids, In: Critical information infrastructures security: 10th international conference, CRITIS 2015, Berlin, Germany, October 5–7, 2015, Revised Selected Papers 10, Springer, pp 199–210

Baker M, Fard AY, Althuwaini H, Shadmand MB (2023) Real-time ai-based anomaly detection and classification in power electronics dominated grids. IEEE J Emerg Select Top Ind Electron 4(2):549–559. https://doi.org/10.1109/JESTIE.2022.3227005

Bauknecht D, Funcke S, Vogel M (2020) Is small beautiful? A framework for assessing decentralised electricity systems. Renew Sustain Energy Rev 118(2019):109543

Bhattacharjee S, Islam MJ, Abedzadeh S (2022) Robust anomaly based attack detection in smart grids under data poisoning attacks, In: Proceedings of the 8th ACM on Cyber-physical system security workshop, pp 3–14

Bruinenberg J, Colton L, Darmois E, Dorn J, Doyle J, Elloumi O, Englert H, Forbes R, Heiles J, Hermans P, Uslar M (2012) CEN -CENELEC-ETSI: smart grid coordination group-smart grid reference architecture report 2.0 (November)

Cali U, Kuzlu M, Pipattanasomporn M, Kempf J, Bai L (2021) Digitalization of power markets and systems using energy informatics. https://doi.org/10.1007/978-3-030-83301-5

Cali U, Kuzlu M, Pipattanasomporn M, Kempf J, Bai L, Cali U, Kuzlu M, Pipattanasomporn M, Kempf J, Bai L (2021) Applications of artificial intelligence in the energy domain. Digitalization of power markets and systems using energy informatics. pp139–168

Choi D-H, Xie L (2017) Impact of power system network topology errors on real-time locational marginal price. J Mod Power Syst Clean Energy 5(5):797–809

De Benedetti M, Leonardi F, Messina F, Santoro C, Vasilakos A (2018) Anomaly detection and predictive maintenance for photovoltaic systems. Neurocomputing 310:59–68

De La Ree J, Centeno V, Thorp JS, Phadke AG (2010) Synchronized phasor measurement applications in power systems. IEEE Trans Smart Grid 1(1):20–27. https://doi.org/10.1109/TSG.2010.2044815

Deng X, Bian D, Wang W, Jiang Z, Yao W, Qiu W, Tong N, Shi D, Liu Y (2020) Deep learning model to detect various synchrophasor data anomalies. IET Gener Trans Distrib 14(24):5739–5745

El Chamie M, Lore KG, Shila DM, Surana A (2018) Physics-based features for anomaly detection in power grids with micro-pmus, In: 2018 IEEE International conference on communications (ICC), IEEE, pp 1–7

Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287–1289

Ford V, Siraj A, Eberle W (2014) Smart grid energy fraud detection using artificial neural networks, In: 2014 IEEE symposium on computational intelligence applications in smart grid (CIASG), IEEE, pp 1–6

Gaggero GB, Rossi M, Girdinio P, Marchese M (2020) Detecting system fault/cyberattack within a photovoltaic system connected to the grid: a neural network-based solution. J Sens Actuator Netw 9(2):20

Gaggero GB, Caviglia R, Armellin A, Rossi M, Girdinio P, Marchese M (2022) Detecting cyberattacks on electrical storage systems through neural network based anomaly detection algorithm. Sensors 22(10):3933

Garza LF, Mandal P (2022) Lstm based hybrid neural network for pmu data forecasting and anomaly detection, In: 2022 North American Power Symposium (NAPS), IEEE, pp 1–6

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press (**http://www.deeplearningbook.org**)

Halden U, Cali U, Catak FO, D'Arco S, Bilendo F (2022) Anomaly detection in power markets and systems. https://arxiv.org/abs/2212.02182

Halden U, Cali U, Catak FO, D'Arco S, Bilendo F (2023) Anomaly detection in power markets and systems, In: 2023 IEEE Power & Energy Society General Meeting (PESGM), IEEE, pp 1–5

Henriksen E, Halden U, Kuzlu M, Cali U (2022) Electrical load forecasting utilizing an explainable artificial intelligence (xai) tool on Norwegian residential buildings, In: 2022 international conference on smart energy systems and technologies (SEST), pp 1–6. https://doi.org/10.1109/SEST53650.2022.9898500

Himeur Y, Ghanem K, Alsalemi A, Bensaali F, Amira A (2021) Artificial intelligence based anomaly detection of energy consumption in buildings: a review, current trends and new perspectives. Appl Energy 287:116601

Hink RCB, Beaver JM, Buckner MA, Morris T, Adhikari U, Pan S (2014) Machine learning for power system disturbance and cyber-attack discrimination, In: 2014 7th international symposium on resilient control systems (ISRCS), IEEE, pp 1–8

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv:1503.02531

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Huang X, Kroening D, Ruan W, Sharp J, Sun Y, Thamo E, Wu M, Yi X (2020) A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. Comput Sci Rev

Huang H, Davis CM, Davis KR (2021) Real-time power system simulation with hardware devices through dnp3 in cyber-physical testbed. IEEE Texas Power Energy Conf 2021:1–6. https://doi.org/10.1109/TPEC51183.2021.9384947

Hundman K, Constantinou V, Laporte C, Colwell I, Soderstrom T (2018) Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding, In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 387–395

Ieee recommended practice for monitoring electric power quality (1995) IEEE Std 1159–1995:1–80. https://doi.org/10.1109/IEEESTD.1995.79050

Jafarnia-Jahromi A, Broumandan A, Nielsen J, Lachapelle G (2012) GPS vulnerability to spoofing threats and a review of antispoofing techniques. Int J Navig Observ. https://doi.org/10.1155/2012/127072

Jamei M, Scaglione A, Roberts C, Stewart E, Peisert S, McParland C, McEachern A (2017) Anomaly detection using optimally placed μPMU sensors in distribution grids. IEEE Trans Power Syst 33(4):3611–3623

Jimada-Ojuolape B, Teh J (2020) Surveys on the reliability impacts of power system cyber-physical layers. Sustain Cities Soc 62:102384

Karney DH (2019) Electricity market deregulation and environmental regulation: evidence from U.S. nuclear power. Energy Econ 84:104500. https://doi.org/10.1016/j.eneco.2019.104500

Karpilow A, Cherkaoui R, D'Arco S, Duong TD (2020) Detection of Bad PMU Data using Machine Learning Techniques, In. IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT) 2020:1–5. https://doi.org/10.1109/ISGT45199.2020.9087782

Krishna VB, Gunter CA, Sanders WH (2018) Evaluating detectors on optimal attack vectors that enable electricity theft and der fraud. IEEE J Select Top Signal Process 12(4):790–805

Kurakin A, Goodfellow I, Bengio S (2016) Adversarial machine learning at scale. arXiv:1611.01236

Lawal OA, Teh J (2023) A framework for modelling the reliability of dynamic line rating operations in a cyber-physical power system network. Sustain Energy Grids Netw 35:101140

Lawal OA, Teh J, Alharbi B, Lai C-M (2024) Data-driven learning-based classification model for mitigating false data injection attacks on dynamic line rating systems. Sustain Energy Grids Netw 38:101347

Noh S-H (2021) Analysis of gradient vanishing of rnns and performance comparison. Information 12(11):442

Ogu RE, Ikerionwu CI, Ayogu II (2021) Leveraging artificial intelligence of things for anomaly detection in advanced metering infrastructures, In: 2020 IEEE 2nd international conference on cyberspac (CYBER NIGERIA), pp 16–20. https://doi.org/10.1109/CYBERNIGERIA51635.2021.9428792

O'Toole Z, Moya C, Rubin C, Schnabel A, Wang J (2019) A cyber-physical testbed design for the electric power grid, In. N Am Power Symp 2019:1–5. https://doi.org/10.1109/NAPS46351.2019.9000312

Ozgur Catak F, Sivaslioglu S, Sahinbas K (2020) A generative model based adversarial security of deep learning and linear classifier models. 2010.08546

Pal S, Sikdar B (2014) A mechanism for detecting data manipulation attacks on pmu data, In: 2014 IEEE international conference on communication systems, IEEE, pp 253–257

Papernot N, McDaniel P, Wu X, Jha S, Swami A (2016) Distillation as a defense to adversarial perturbations against deep neural networks. arXiv:1511.04508

Pardha Saradhi J, Srinivasarao R, Ganesh V (2020) Wavelet based multiresolution analysis of a 5-Bus system in the presence SVC controller under fault and sudden load conditions, Mater Today. https://doi.org/10.1016/j.matpr.2020.10.852https://www.sciencedirect.com/science/article/pii/S2214785320384893

Phadke AG, Bi T (2018) Phasor measurement units, wams, and their applications in protection and control of power systems. J Mod Power Syst Clean Energy 6(4):619–629. https://doi.org/10.1007/s40565-018-0423-3

Pukelsheim F (1994) The three sigma rule. Am Stat 48(2):88–91

Qayyum A, Usama M, Qadir J, Al-Fuqaha A (2020) Securing connected autonomous vehicles: challenges posed by adversarial machine learning and the way forward. IEEE Commun Surv Tutor 22(2):998–1026. https://doi.org/10.1109/COMST.2020.2975048

Rafferty M, Brogan P, Hastings J, Laverty D, Liu XA, Khan R (2018) Local anomaly detection by application of regression analysis on pmu data, In: 2018 IEEE Power & Energy Society General Meeting (PESGM), IEEE, pp 1–5

Ramasubramanian B, Rajan MA, Girish Chandra M, Cleaveland R, Marcus SI (2022) Resilience to denial-of-service and integrity attacks: a structured systems approach. Eur J Control 63:61–69. https://doi.org/10.1016/j.ejcon.2021.09.005

Ren H, Hou Z, Etingov P (2018) Online anomaly detection using machine learning and hpc for power system synchrophasor measurements, In. IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS) 2018:1–5. https://doi.org/10.1109/PMAPS.2018.8440495

Risbud P, Gatsis N, Taha A (2019) Vulnerability analysis of smart grids to GPS spoofing. IEEE Trans Smart Grid 10(4):3535–3548. https://doi.org/10.1109/TSG.2018.2830118

Roy P, Bhattacharjee S, Das SK (2020) Real time stream mining based attack detection in distribution level pmus for smart grids, In: GLOBECOM 2020-2020 IEEE global communications conference, IEEE, pp 1–6

Sadeghi K, Banerjee A, Gupta SKS (2020) A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE Trans Emerg Top Comput Intell 4(4):450–467. https://doi.org/10.1109/TETCI.2020.2968933

Sharadga H, Hajimirza S, Balog RS (2020) Time series forecasting of solar power generation for large-scale photovoltaic plants. Renew Energy 150:797–807. https://doi.org/10.1016/j.renene.2019.12.131

Sivasankari N, Kamalakkannan S (2022) Detection and prevention of man-in-the-middle attack in iot network using regression modeling. Adv Eng Softw 169

Styvaktakis E, Gu IYH, Bollen MHJ (2003) Event-based transient categorization and analysis in electric power systems, In: SMC'03 Conference Proceedings. 2003 IEEE international conference on systems, man and cybernetics. Conference theme-system security and assurance (Cat. No.03CH37483), vol 5, pp. 4176–4183. https://doi.org/10.1109/ICSMC.2003.1245641

Summary for Policymakers—Global Warming of 1.5 °C. https://www.ipcc.ch/sr15/chapter/spm/

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. arXiv:1312.6199

Tinawi I (2019) Machine learning for time series anomaly detection, Ph.D. thesis, Massachusetts Institute of Technology

Tu C, He X, Liu X, Li P (2018) Cyber-attacks in pmu-based power network and countermeasures. IEEE Access 6:65594–65603

Uslar M, Delfs C, Gottschalk M (2017) The IEC 62559-2 Use Case Template and the SGAM Applied in Various Domains

Valdes A, Macwan R, Backes M (2016) Anomaly detection in electrical substation circuits via unsupervised machine learning, In: 2016 IEEE 17th international conference on information reuse and integration (IRI), IEEE, pp 500–505

Veerakumar N, Ćetenović D, Kongurai K, Popov M, Jongepier A, Terzija V (2023) PMU-based real-time distribution system state estimation considering anomaly detection, discrimination and identification. Int J Electr Power Energy Syst 148:108916

Vicol B, Gavrilas M, Ivanov O (2013) Modern Technologies for Power Systems Monitoring, ELS International Symposium (June)

Wang J, Shi D, Li Y, Chen J, Ding H, Duan X (2018) Distributed framework for detecting pmu data manipulation attacks with deep autoencoders. IEEE Trans Smart Grid 10(4):4401–4410

Yang Z, Chen N, Chen Y, Zhou N (2018) A novel pmu fog based early anomaly detection for an efficient wide area pmu network, In: 2018 IEEE 2nd International Conference on Fog and Edge Computing (ICFEC), IEEE, pp 1–10

Zhou M, Wang Y, Srivastava AK, Wu Y, Banerjee P (2018) Ensemble-based algorithm for synchrophasor data anomaly detection. IEEE Trans Smart Grid 10(3):2979–2988

Zhou Y, Arghandeh R, Konstantakopoulos I, Abdullah S, von Meier A, Spanos CJ (2016) Abnormal event detection with high resolution micro-pmu data, In: 2016 Power Systems Computation Conference (PSCC), IEEE, pp 1–7

Authors and Affiliations

Umit Cali[1,3] · Ferhat Ozgur Catak[2] · Ugur Halden[3]

✉ Umit Cali
umit.cali@ntnu.no

Ferhat Ozgur Catak
f.ozgur.catak@uis.no

Ugur Halden
ugur.halden@ntnu.no

[1] School of Physics, Engineering and Technology, University of York, Heslington, York YO10 5DD, UK

[2] The Faculty of Science and Technology, Department of Electrical Engineering and Computer Science, University of Stavanger, Kjell Arholms gate 41, 4021 Stavanger, Norway

[3] Department of Electric Energy, Norwegian University of Science and Technology, O. S. Bragstads plass 2E, 7034 Trondheim, Norway