Enhancing Situational Awareness in Low-Voltage Grids through Digital Process Twins and Data-Driven Methods

A thesis approved for the academic degree of

Doctor of Engineering (Dr.-Ing.)

at the

Faculty of Electrical Engineering
and Information Technology

TU Dortmund University

by

Razieh Balouchi Anaraki

June 2025

Supervisor: Professor Dr.-Ing. Christian Rehtanz

Co-Advisor: Professor Dr.-Ing. Wolfram Wellßow

Day of Oral Examination: 26.01.2026

Abstract

The low-voltage power grid is experiencing significant transformations due to the integration of renewable energy sources and the introduction of new types of loads, including electric vehicles, electric heat pumps, and smart home systems. These developments add complexity and variability, posing significant challenges to conventional grid management methods. As a result, there is a growing need for digitalization, particularly in low-voltage grids, to support timely and informed decision-making by distribution system operators.

In response, the implementation of digital technologies, particularly digital process twins, has emerged as a promising solution. Digital process twins provide a model of grid operations, capturing both manual and automated tasks. Digital process twins provide real-time simulation and analysis of the power grid, enabling operators to visualize and manage the grid's behavior effectively. By leveraging real-time data analysis and artificial intelligence, digital process twins facilitate a deeper understanding of grid operations, enhancing the ability to preempt potential issues and thus improving the grid’s stability and reliability.

The core objective of this thesis is to enhance situational awareness within low-voltage grids by utilizing the advanced capabilities of digital process twins. This involves employing sophisticated data-driven methodologies to forecast potential critical scenarios and detect anomalies that could disrupt grid functionality, capabilities that are essential for adapting to new consumer behaviors, managing potential grid failures, and accommodating the variable nature of renewable energy sources.

Moreover, this research addresses the growing complexity and operational demands of modern power systems, which often challenge distribution system operators' ability to maintain a comprehensive and accurate understanding of the grid state. These challenges can hinder their ability to attain the necessary level of situational awareness required to make informed decisions and respond effectively to incidents.

This thesis aims to demonstrate how digital process twins can be strategically utilized to improve situational awareness in the low-voltage grid. By utilizing data and artificial intelligence-driven insights, this research aims to establish a foundation for a more resilient and efficient power system, ensuring the grid is well-equipped to meet the demands of an increasingly complex and rapidly evolving energy environment.

Kurzfassung

Das Niederspannungsnetz erfährt durch die Integration erneuerbarer Energiequellen und die Einführung neuer Arten von Lasten wie Elektrofahrzeuge, elektrische Wärmepumpen und Smart-Home-Systeme transformative Veränderungen. Diese Entwicklungen führen zu mehr Komplexität und Variabilität und stellen herkömmliche Netzmanagementmethoden vor erhebliche Herausforderungen. Als Reaktion darauf ist die Implementierung digitaler Technologien, insbesondere digitale Prozesszwillinge, von entscheidender Bedeutung. Digitaler Prozesszwillinge in Stromnetzen bieten ein digitales Modell der von Netzbetreibern durchgeführten betrieblichen Aufgaben, das sowohl manuelle technische Aufgaben als auch solche umfasst, die von Software ausgeführt werden.

Digitale Prozesszwillinge bieten Echtzeitsimulationen und -analysen des Stromnetzes und ermöglichen es den Betreibern, das Verhalten des Netzes effektiv zu visualisieren und zu steuern. Durch den Einsatz von Echtzeit-Datenanalyse und künstlicher Intelligenz ermöglichen digitaler Prozesszwillinge ein tieferes Verständnis des Netzbetriebs, was die Fähigkeit zur Vorbeugung potenzieller Probleme verbessert und somit die Stabilität und Zuverlässigkeit des Netzes erhöht.

Das Hauptziel dieser Arbeit ist die Verbesserung der Situation in Niederspannungsnetzen durch die Nutzung der fortschrittlichen Fähigkeiten digitaler Prozesszwillinge. Dies beinhaltet den Einsatz hochentwickelter datengesteuerter Methoden zur Vorhersage potenziell kritischer Szenarien und zur Erkennung von Anomalien, die die Netzfunktionalität stören könnten, Fähigkeiten, die für die Anpassung an neue Verbraucherverhaltensweisen, die Bewältigung potenzieller Netzausfälle und die Berücksichtigung der variablen Natur der erneuerbaren Energien unerlässlich sind.

Darüber hinaus befasst sich diese Forschung mit der zunehmenden Komplexität und den betrieblichen Anforderungen moderner Stromnetze, die die Fähigkeit von Verteilnetzbetreibern, ein umfassendes und genaues Verständnis des Netzzustands zu erhalten, oft in Frage stellen. Diese Herausforderungen können ihre Fähigkeit beeinträchtigen, das erforderliche Maß an Sicherheit zu erreichen, um fundierte Entscheidungen zu treffen und wirksam auf Zwischenfälle zu reagieren.

Diese Arbeit soll zeigen, wie digitaler Prozesszwillinge strategisch eingesetzt werden können, um die Sicherheit im Niederspannungsnetz zu verbessern. Durch die Nutzung von Daten und künstlicher Intelligenz-gesteuerten Erkenntnissen soll diese Forschung eine Grundlage für ein widerstandsfähigeres und effizienteres Stromsystem schaffen und sicherstellen, dass das Netz gut auf die Anforderungen einer zunehmend komplexen und sich schnell entwickelnden Energieumgebung vorbereitet ist.

Acknowledgments

First and foremost, I extend my deepest gratitude to my advisor, Professor Dr.-Ing. Christian Rehtanz, for his invaluable guidance and unwavering support throughout this thesis. His mentorship not only shaped this research but also enriched my personal and professional growth during my time at the university.

I am profoundly thankful to Professor Dr.-Ing. Wolfram Wellßow, whose trust and invitation allowed me to migrate to Germany and embark on this significant and transformative journey in my life. His belief in my potential has been a great encouragement.

I would also like to express my appreciation to the members of my committee for generously offering their time and guidance.

Special thanks are due to Professor Dr.-Ing. Ulf Häger for his support and thoughtful feedback throughout this research.

I would like to thank my family for their endless love, care, support, and motivation, without which I would not have been able to achieve my goal.

Last but not least, many thanks go to all my colleagues at the ie^3 group at TU Dortmund and the ESEM group at RPTU (TU Kaiserslautern) for the good times during my doctoral studies.

This thesis is not only a reflection of my work but also the support of each of these individuals. Thank you for making this journey rewarding and successful.

Bayreuth, June 2025

Razieh Balouchi Anaraki

Contents

1 Introduction

This chapter presents the motivation for this research, emphasizing the increasing complexity of Low-Voltage (LV) grids. It introduces the concept of the Digital Process Twin (DPT), which extends the Digital Twin (DT) by including not only physical system models but also operator tasks and real-time analytics. Within this framework, situational awareness (SA) is identified as part of DPT to ensure secure and efficient grid operation. While several methods exist to improve SA, this thesis focuses specifically on two key approaches: forecasting and anomaly detection, which are explored in detail in the following chapters.

1.1 Research Motivation

The integration of distributed energy resources, such as solar, wind, and storage, is transforming electricity distribution [2]. This shift toward decentralized energy reduces grid losses by locating generation nearer to demand centers and lowers carbon emissions through cleaner, localized production [3]. However, the variability and decentralization of these resources pose significant forecasting challenges due to their limited predictability [4]. Transitioning from centralized to distributed generation involves managing bidirectional power flows and coping with the increased volatility from renewable sources. This requires sophisticated demand flexibility and management strategies to maintain grid stability [5].

Additionally, the expansion of smart grid technologies, such as advanced metering infrastructure, has significantly enhanced real-time electricity monitoring and control capabilities. However, this influx of data also adds complexity and strains traditional grid management methods, which often cannot keep up with the dynamic nature of modern energy flows. This challenge is further complicated by the variable availability of renewable energy sources, which can result in unpredictable grid behaviors, especially under adverse weather conditions. These conditions have been shown to significantly impact power systems, often leading to sudden and severe disruptions [6].

Effective SA in electric grids involves understanding the grid's current environment and predicting changes, notably due to factors such as cybersecurity threats [7]. Achieving effective SA is crucial for proactive grid management and customer satisfaction, as it enables operators to monitor and anticipate system conditions, reduce outage durations, and facilitate quicker restorations, thus minimizing economic losses [8]. Additionally, it supports strategic decision-making in load balancing and infrastructure maintenance, enhancing grid reliability and reducing the likelihood of failures.

SA is particularly crucial in LV grids due to unique challenges that compound traditional issues. These include [9]:

- Radial or weakly meshed grids: LV grids are essentially built in a radial topology, which

is simple and cost-effective to implement; however, it lacks structural redundancy. As a result, a single fault can disrupt the power flow to any point beyond the fault [10].

- Poor observability: There is a lack of comprehensive monitoring, especially due to the limited installation of real-time meters. In distribution grids, especially at the medium voltage and LV levels, a primary challenge is the lack of adequate monitoring at key locations, such as exits from transformer stations. This issue results in significant visibility problems within the deeper layers of the grid [11].
- Unbalanced operation: The operational imbalance due to varying loads and the existence of single- and two-phase branches complicates power management.
- Dynamic network configurations: The configuration of LV grids often changes without being recorded [12], making the actual network layout 'invisible' and complicating effective management and decision-making [13].

To improve SA in power grid operations, operators need to have systems that can capture, transmit, analyze, and store data. These capabilities allow operators to quickly identify any unusual activity, take corrective actions, investigate what caused these issues, and share their findings across the electrical grid and all aspects of operational planning. The integration of DPT technology, which involves creating a dynamic digital representation of physical systems to simulate real-world conditions and processes, significantly amplifies these capabilities by creating a live digital model of the physical grid. Creating a live digital model of the physical grid can enhance predictive analytics and maintenance strategies, provided that the model accurately represents key system properties and is supported by reliable data inputs [14].

Based on this technology, this thesis proposes the adoption of a DPT as a transformative tool for enhancing SA in the power grid. A DPT provides a real-time virtual model of the grid, equipped with advanced data analytics and Artificial Intelligence (AI)-driven insights. This model mirrors the grid's physical state and simulates various operational scenarios, allowing grid operators to anticipate and react to potential disruptions in a proactive manner.

To improve SA, this thesis integrates anomaly detection and Pseudo-Worst-Case Forecasting (PWCF) into the DPT framework. These key features enhance the grid's resilience by providing assistive signals that help operators manage complex scenarios more effectively. PWCF forecasts worst-case scenarios and undesirable events, leveraging both real-time grid conditions and historical data. By providing these predictive insights, the DPT equips operators with the necessary foresight to implement preventive control strategies, mitigating risks before they materialize.

Anomaly detection is another vital component within the DPT framework. Outliers, observations that significantly deviate from normal behavior [15], are well-documented in research. However, this thesis goes further by categorizing different types of anomalies and

analyzing how their characteristics influence the effectiveness of detection methods. Using advanced data-driven and AI techniques, the DPT continuously monitors the grid, processing data from smart meters, substations, and Internet of Things (IoT) devices to detect deviations from expected patterns. These deviations, or anomalies, may indicate potential issues such as equipment malfunctions, cyberattacks, or unexpected changes in energy consumption. Early detection is critical, as undetected anomalies can escalate, leading to substantial financial losses or, in severe cases, system-wide failures [16].

One specific application of anomaly detection within the DPT framework is the identification of new or unexpected loads. If such loads are energized without prior coordination or system awareness, they may disrupt restoration procedures, overload local grid segments, or even initiate cascading failures. Detecting these loads in advance allows operators to adjust forecasting models and restoration strategies accordingly, thus preventing instability.

Applications of new load detection include:

- Managing power system restoration: Electric heat pumps and electric vehicles, with their high power demand, intensify the cold load pickup effect, increasing the risk of grid overload during restoration [17]. As their adoption grows, careful planning becomes essential to ensure stable and efficient power system recovery.
- Preventing system overload: New loads, like electric heat pumps and electric vehicles, can lead to localized grid overloads or cause cascading failures.
- Improving load forecasting: Detecting and categorizing new loads ensures that load models are updated.

The core components of SA, including its definition, the reasons it is needed in power systems, improvement methods, and expected advantages, are summarized in Figure 1.1.

1.2 Research Gap and Objective

With the growing size and complexity of modern power systems, operators are increasingly facing challenges in obtaining a precise understanding of the systems under their management. These challenges prevent their ability to achieve the necessary level of SA required for making the right decisions and responding effectively to incidents [18]. This section highlights critical gaps in current methodologies used for monitoring LV grids, emphasizing the need for an advanced approach utilizing DPT to improve SA and system reliability.

- Reliance on manual processing and traditional event detection: Traditional event detection in power systems often depends on manual processes, leading to delays and inaccuracies in event detection. This thesis proposes the integration of advanced data driven and AI-based methods within DPT frameworks to automate and enhance the accuracy of these processes.

Definition of SA

‘‘Information gathered from a variety of sources that, when communicated to emergency managers and decision-makers, can form the basis for incident management decision making’’[1].

Factors driving the need for SA in power grids

Increasing complexity and interconnectivity of grid structures

Limited system visibility and insufficient data analysis capabilities

Need for advanced tools and strategies for effective grid monitoring

Methods to improve SA

Deployment of advanced measurement technologies.

Integration of DT and DPT frameworks (SA is a function of the DPT framework, enabling operators to perceive, interpret, and anticipate grid states by integrating data, operator tasks, and real-time analytics).

Application of advanced data analytics and AI-based methods, including:

- State estimation
- Forecasting
- Anomaly detection

Advantages of SA

Comprehensive real-time grid monitoring

Optimized demand management

Rapid response to outages and issues

Improved risk management and mitigation

Enhanced integration and efficiency of distributed energy resources

Figure 1.1: Overview of SA in power grids within the DPT framework

- Inaccuracies in topology-based algorithms: Inaccuracies in network parameters like line admittance and load models often render grids as "black boxes," where precise conditions are unknown. Developing new algorithms that are less dependent on precise data or that can effectively estimate missing parameters is crucial for maintaining accuracy and reliability.
- Data integration issues: Real-time integration of data from loads, renewable sources, and LV systems remains insufficient, despite being essential for emergency operations and system restoration. A comprehensive platform that aggregates data from all relevant sources is imperative to strengthen operators' decision-making.
- Lack of advanced decision support tools: Current systems cannot fully utilize complex data models for predicting system responses, which restricts effective decision-making during emergencies.
- Standardization and cooperation issues: Effective power system restoration is currently hampered by a lack of standardized processes for interaction and data exchange between Transmission System Operators (TSOs) and Distribution System Operators (DSOs). Establishing standardized protocols and enhancing cooperative operations are essential for leveraging renewable energy sources more efficiently and improving overall system responsiveness.
- Suboptimal utilization of real-time data: The predominant use of offline databases limits dynamic SA and real-time control capabilities during critical operational scenarios. Transitioning to a DT that supports real-time data processing will provide immediate and actionable insights, enhancing SA and enabling better control during power system restoration.

This gap in monitoring is compounded by the frequent, daily changes in grid conditions driven by load fluctuations, reverse power flows, and potential bottlenecks. Traditional local network substations often suffer from a lack of remote controllability, which introduces delays and inefficiencies in managing switching states, thereby affecting the overall responsiveness of the network management.

To address these issues, the adoption of fully digital, remotely monitored, and controlled local network substations is proposed. DT and DPT provide innovative solutions to these issues by enabling:

- Enhanced SA: DPT simulates the grid's physical state in real time, offering continuous monitoring and diagnostics that improve responsiveness to disruptions.
- Full digitalization: This feature allows for comprehensive state diagnostics, extending maintenance intervals, precise measurements at network node points to validate and

extrapolate network calculations without recalculation, and fully digitized controls. AI technology, which is gaining mainstream attention, underpins these capabilities [19].

- Network automation and remote reporting: Automated systems enhance grid management by precisely identifying errors and enabling remote switch operations, paving the way for advanced functionalities such as programmability, reducing the need for manual intervention, and boosting operational efficiency.
- Data distribution: By facilitating standardized data exchange and operational processes, DT and DPT improve coordination across the grid, ensuring scalability and adaptability to future demands.

This thesis explores the core concepts and practical applications of DPT in enhancing SA within LV grids, with a focus on forecasting and anomaly detection. Notably, Machine Learning (ML) methods stand out as the most prevalent technique in anomaly detection, utilized in approximately 23.3% of power system studies. In power generation contexts specifically, their application rises to 30.9%. Despite their prominence, a significant research gap exists in the application of ML for detecting anomalies in consumer-side operations, particularly within smart grids and energy consumption. Additionally, the application of DT and data analytics in this field accounts for only 2.7% [20]. This indicates substantial potential for integrating DT and ML to develop comprehensive strategies for anomaly detection on the consumer side. To address the research gaps in digitalization, predictive grid management, and AI-driven operational enhancements, the following research questions will be explored:

1. Why are digitalization and DT essential for the operational enhancement of power systems?
2. How does the application of DPT contribute to SA in power systems?
3. What improvements can be made to SA to address the challenges faced by modern power systems and LV grids?
4. How can DPT provide assistive signals that help operators make better decisions?

1.3 Dissertation Outline

The organization of this thesis is as follows: Chapter 1 introduces the research by outlining the motivation behind the study, identifying existing gaps, and articulating the objectives aimed at bridging these gaps. Chapter 2 delves into the DT technology, tracing its evolution in the digitization of power grids, discussing various architectures, and detailing its application in power systems, particularly focusing on the complexities of data and the challenges associated with big data in smart grids. Chapter 3 focuses on enhancing SA in power systems, defining it specifically for distribution grids and examining methods such as forecasting and anomaly

detection to improve it.

Chapter 4 describes the methodology used in this research, including data preprocessing and analysis, and introduces the PWCF. It elaborates on different forecasting techniques, including advanced methods integrating Neural Networks (NN), and explores various ML approaches for anomaly detection. Chapter 5 presents case studies and simulation results, providing a detailed overview of the datasets employed and discussing the effectiveness of the methods used, such as PWCF and various anomaly detection techniques.

Chapter 6 concludes the thesis by summarizing the key findings, discussing the implications of the research, and outlining directions for future work.

2 Digital Twin Technology for Power Systems

As power systems grow more complex due to the increasing integration of distributed energy resources and new loads, traditional monitoring and control methods fall short. This chapter introduces the need for digitalization in power systems as a response to these growing challenges. It presents Digital Twin (DT) and Digital Process Twin (DPT) technologies as key enablers of real-time monitoring, simulation, and decision support. While DT focuses on virtual representations of physical assets, DPT extends this by integrating system processes, operator tasks, and data analytics. The chapter also outlines the advantages of DT and DPT, such as improved system visibility, predictive maintenance, and operational optimization, as well as the technical and data-related challenges they face. This thesis focuses specifically on the knowledge extraction layer of the DPT, which transforms real-time data into actionable signals to improve Situational Awareness (SA).

2.1 History of Power Grid Digitization

As power systems evolve to meet the increasing demand for renewable energy sources and the emergence of new types of loads, they face growing complexity across all voltage levels, from generation and transmission to distribution, which necessitates the use of advanced monitoring, control, and planning strategies. These challenges affect the entire grid infrastructure but are particularly evident at the distribution level, where technologies such as photovoltaic (PV) systems, electric vehicles, and electric heat pumps are increasingly connected. As a result, the grid is becoming more complex and heterogeneous, with nonlinear dynamics that make it harder to model and understand [21]. Digitalization becomes essential, and provides the tools necessary to optimize grid performance, integrate renewable energy sources, and enhance overall grid visibility [22, 23]. A pioneering power grid is characterized by an emphasis on information, digital transformation, automation, and interaction [24].

A significant step towards digitalization occurred in 2007 with the passage of the Energy Independence and Security Act[1], which established a national policy recognizing the National Institute of Standards and Technology[2] for leadership in modernizing the electric grid [25]. Moreover, integrating phasor measurement units into future control systems has become crucial for effectively monitoring, observing, and analyzing power system dynamics, and adds an independent observability layer separate from the Supervisory Control and Data Acquisition (SCADA) system, which allows for the monitoring and analysis of dynamic events, even if the SCADA system is not functioning properly [26]. These advancements underscore the need for continuous innovation in power grid monitoring strategies. Despite the potential of emerging technologies, the pace of grid digitalization remains uneven. For example, smart meters,

[1] EISA
[2] NIST

primarily deployed at the consumer level, have achieved penetration rates exceeding 80% in countries such as Italy, Denmark, and Estonia, whereas in others, including Germany, adoption remains below 20% [27]. This uneven rollout highlights the challenges in realizing the full benefits of grid digitalization and smart meters represent only one aspect of digitalization.

The digital transformation of power systems has progressed through distinct technological milestones, each expanding capacity for automation, observability, and intelligence. Initially driven by Operational Technology (OT), such as SCADA systems in the 1960s, digitalization focused on enabling centralized control and remote supervision of physical infrastructure. Later, the emergence of smart grid technologies introduced greater integration of Information Technology (IT) functions, including data analytics, cloud platforms, and bidirectional communication between grid operators and consumers. In recent years, the development of DT has marked a phase of full IT/OT convergence, where real-time modeling, AI-based forecasting, and cyber-physical synchronization provide predictive insights and automation at scale. Table 2.1 shows the history of power grid digitalization, highlighting milestones in the transition toward digitized power systems.

Table 2.1: Key Milestones in the digitization of power systems

Technology	Periode	Description
Traditional power system	1880	Edison's Direct Current (DC) generating station at 257 Pearl Street began supplying electricity to customers.
SCADA (OT)	early 1960s	Supervisory control refers to a high-level system that oversees and coordinates multiple individual controllers or control loops.
Wide area monitoring system (OT+ IT)	Late 1980	It consists of advanced measurement technology, information tools, and operational infrastructure that enable the planning, operation, and management of large and complex electric power systems. It is expressly designed to enhance the operator's real-time SA [28].
Smart grids (IT+ OT)	21th century	It can communicate, store data, and make decisions based on analysis. It improves traditional grids by enabling faster, more efficient, and integrated operations[29].
DT (Full IT/OT convergence)	2020s	A physical grid that is highly interconnected and decentralized, becoming smarter through integration with advanced communication technologies (e.g. 4G/5G), distributed ledgers, and AI [30].

2.2 An Introduction to Digital Twin Technology

DT is a virtual representation of an existing or future physical object or system. This virtual model incorporates descriptions of its attributes and functional properties and maintains a live link with its real-world counterpart. This link facilitates continuous updates and adjustments through digital autonomous communication infrastructures or manual interventions, ensuring that the twin evolves alongside its physical counterpart from inception through disposal [31]. The concept of DT has gained significant traction over the past decade, especially within industries that demand high accuracy and real-time updates, such as manufacturing, aviation, and energy. The journey of DT technology began in the early 1960s with the use of online digital computers in the power industry to improve boiler turbine efficiency and increase safety through advanced data management [32]. The National Aeronautics and Space Administration (NASA) used digital technologies during its space exploration as a "living model" of the Apollo mission in the 1960s [33]. In 1991, David Gelernter, in his book Mirror World, conceptualized a parallel digital world that mirrors physical reality in detail [34]. In 2002, Dr. Michael Graves at the University of Michigan applied the DT concept to manufacturing and demonstrated its practical applications beyond theoretical ideas [35]. The term DT was popularized by NASA by John Vickers in 2010, who emphasized its use in space exploration and its ability to simulate, predict, and optimize complex systems [36]. The concept of Industry 4.0, introduced by the German federal government as part of its high-tech strategy in 2011, has influenced the creation of a new simulation modeling paradigm embodied by the DT concept [37].

Figure 2.1 illustrates the historical progression of DT technology from early use in power industries to Industry 4.0.

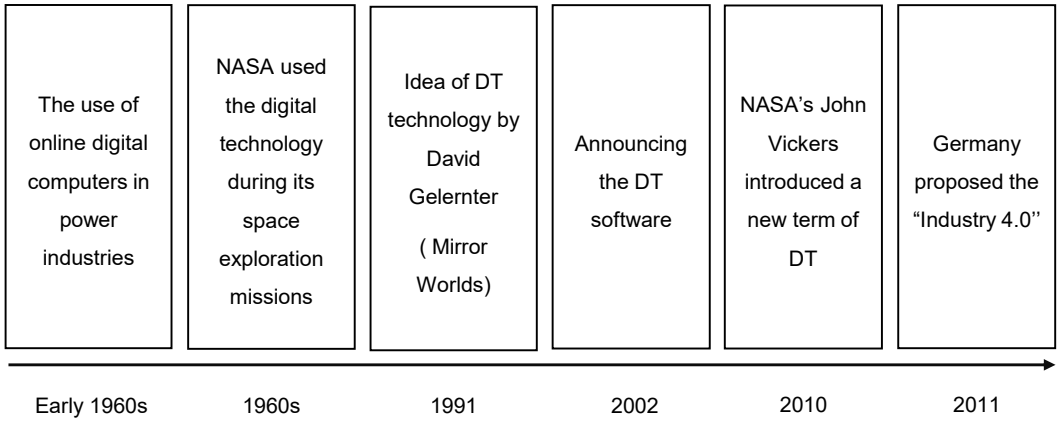


Figure 2.1: The historical progression of DT technology

As illustrated in Table 2.2, the concept of DTs has evolved, with various definitions emerging from different experts and sources. Table 2.2 provides a summary of some key definitions from notable authors and publications, showcasing the diverse perspectives and applications of DT technology across different fields and industries.

Table 2.2: Definition of DT in different sources

Reference	DT Definition
NASA 2012 [38]	"A DT is an integrated multiphysics, multiscale, probabilistic simulation of an as-built vehicle or system that uses the best available physical models, sensor updates, fleet history, etc, to mirror the life of its corresponding flying twin.''
Rosen et al. 2015 [39]	"The autonomous systems will need access to very realistic models of the current state of the process and their behavior in interaction with their environment in the real world, typically called the DT.''
Bochert et al. 2016 [40]	"The DT refers to a comprehensive physical and functional description of a component, product, or system, which includes more or less all information that could be useful in all the current and subsequent lifecycle phases.''
Grieves et al. 2017 [41]	"The DT is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the microatomic level to the macro geometrical level.''
Liu et al. 2018 [42]	"The DT is a living model of the physical asset or system, which continually adapts to operational changes based on the collected online data and information, and can forecast the future of the corresponding physical counterpart.''

DTs can represent objects, systems, or functions, serving as crucial tools from the design stage through to manufacturing. By identifying potential issues and risks early in the design process, DTs support improvements in the final design. Their utility extends beyond the initial stages into commissioning, operation, and disposal, thereby encapsulating the entire lifecycle of a product. Unlike conventional simulation tools such as Finite Element (FE) or Electromagnetic Transient (EMT) models, which are domain-specific and scenario-based, DTs provide a holistic and continuously updated representation of the physical system by integrating real-time data. Through this capability, DTs can learn from past data, adapt to present conditions, and forecast future scenarios, thereby enhancing decision-making at every stage. DTs are employed across multiple domains, including healthcare, industry, aviation, and energy [43]. Table 2.3 highlights different use cases of DT technology, with a particular emphasis on a few specific examples.

Table 2.3: Applications of DT technology across different fields

Field of study	Example
Power industry	Early work in [32] introduced online digital computers and digital modeling for power systems in the US. While not yet a full DT, this effort marked an important precursor by combining simulation models with operational data, paving the way for later DT applications
Smart cities	The referenced study first introduced the DT of the city of Zurich [45].
Manufacturing	A fault diagnosis method that utilizes DT technology to achieve intelligent manufacturing was introduced in [46].
Healthcare	DT was used to create virtual reality in developing a physical robot fish, as detailed in [47].
Energy	A tool for energy analysis aimed at improving energy efficiency was introduced in [48].

As illustrated in the pie chart in Figure 2.2, based on a report published on June 3, 2024 [44], over 50 percent of all research related to DT focuses on manufacturing. The energy sector, with 17 percent, ranks second in terms of DT application research. Within the energy sector, DT technology application is primarily employed for performance monitoring, fault detection, enhancing power plant efficiency, and facilitating the expansion and development of energy grids. This data highlights the increasing importance of DT in enhancing the operational efficiencies and capabilities of various sectors, extending beyond their initial industrial applications.

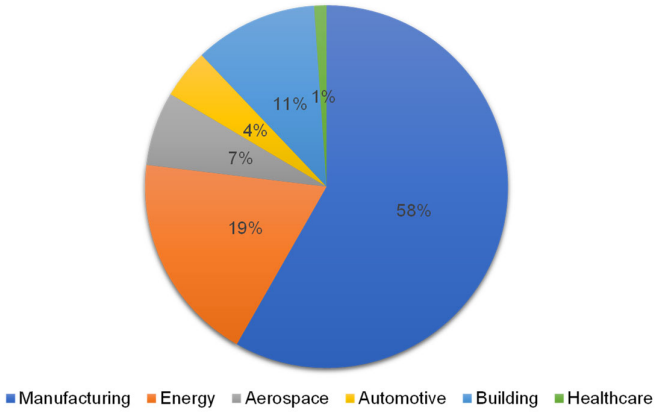


Figure 2.2: Distribution of main application areas of DT

2.3 Architectures and Challenges of Digital Twins

DTs are categorized into several types based on their application goals [49], as shown in Figure 2.3. The architecture of DTs begins with the smallest functional unit, known as

component twins. These are individual elements of a larger system. When two or more components are integrated, they form what is known as an asset. These assets comprise various components that function collectively. The system twins show the formation of a complete system by various assets coming together. The twin process shows how the systems combine to achieve collective functionality. At a higher level, system twins represent the merger of multiple assets. The integration and interaction among these systems are further explored through process twins, which illustrate how various systems collaborate within a full production facility.

A DT can incorporate a variety of digital models to enhance its predictive and operational capabilities [50], including:

- Numerical model: This model can rely on data-driven methods and uses AI algorithms.
- 1-D simulation: A flow diagram is paired with blocks that simulate the system's performance in this model.
- 3-D simulation: This model is used to depict the dynamics and structure of the system.

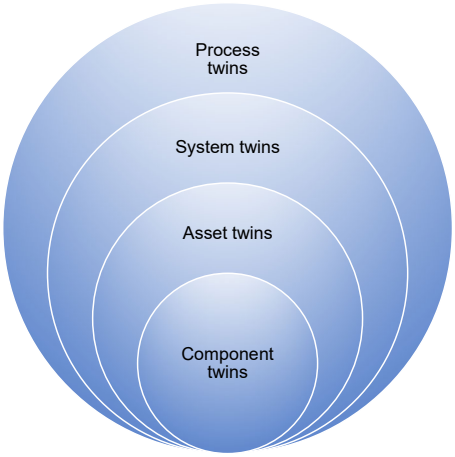


Figure 2.3: Different types of DT based on the area of applications

Understanding the levels of integration of DTs is crucial for comprehending the connectivity between the physical and digital components, which is a fundamental aspect that impacts their functionality and application. These integration levels can be described as follows, and are illustrated in Figure 2.4 [51–53].

- Digital model: This is the basic level of integration, where manual updates are necessary to align the digital representation with its physical counterpart. There is no automatic data exchange between the digital and physical components, making this the least integrated form. It can be a simple or detailed representation designed for analysis, simulation, or visualization. Depending on the context, digital models can exist in different forms, including 3D models, mathematical models, or data models. For example, in power grids, this model represents the design and configuration of the grid,

including its components like substations, transmission lines, and power generation facilities.

- Digital generator: A digital generator is seen as a system or component that produces digital data or simulations. It typically refers to a software tool or algorithm that creates digital representations of entities or processes.
- Digital shadow: This level involves a unidirectional automatic data flow from the physical world to the digital model. Information is seamlessly transferred to the digital environment, but modifications in the digital realm do not affect the physical entity. The purpose of a digital shadow is to provide insight into the current state or behavior of the corresponding physical entity. For example, in power grids, a digital shadow receives real-time data from the power grid's physical components via sensors and provides operators with up-to-date information on system performance, energy flow, and potential issues.
- DT: A Digital Twin represents a dynamic, continuously updated digital replica of a physical system or component. It not only exchanges data bidirectionally with its physical counterpart but also simulates, predicts, and analyzes its behavior in real time.

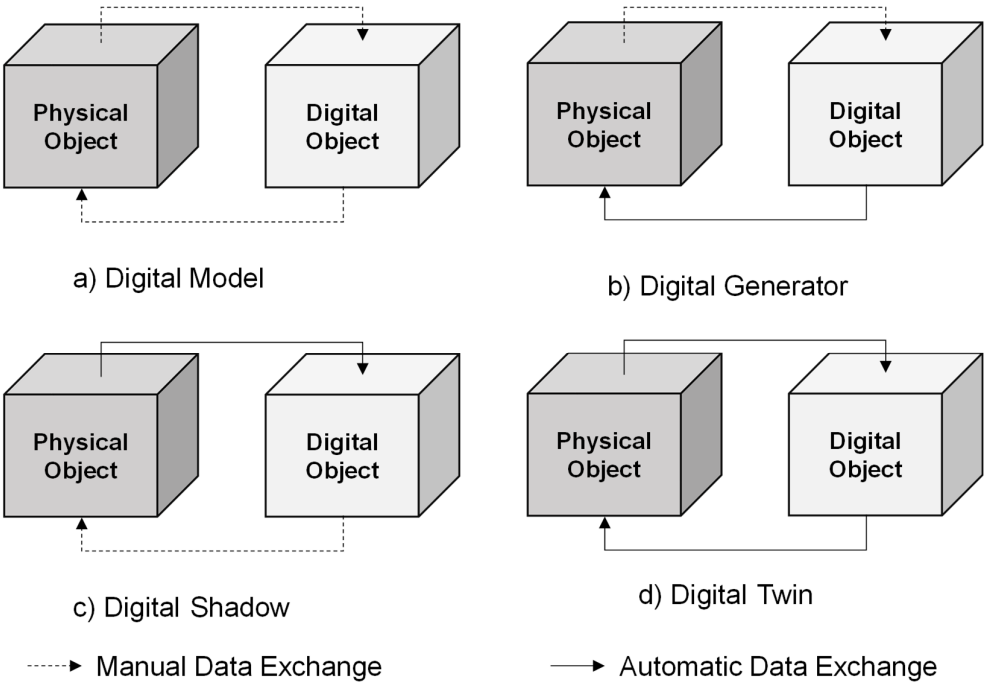


Figure 2.4: Difference between different digital components (a) digital model, (b) digital generator, (c) digital shadow, and (d) DT

The creation of effective DTs necessitates robust system models, which in turn require precise system identification. To be able to perform system identification, data-driven methods are relied upon. Advances in the IoT and AI have facilitated this identification process by providing

extensive data about system components and behaviors. The evolution of DT technology parallels that of AI and the IoT, presenting common challenges that affect their development and efficiency. Challenges in DTs, AI, and IoT are multifaceted and often interconnected. These challenges are detailed in Table 2.4 [43].

Table 2.4: Challenges of DT technology

AI challenges	
IT infrastructure	The costs of high-performance graphics processing units capable of performing machine learning and deep learning algorithms
Data	Ensuring that the AI algorithms are fed with the highest quality data
Privacy and security	Regulations and measures concerning AI will need to be developed as the technology continues to grow
IoT challenges	
Data, privacy, security	Collecting a substantial amount of data Controlling the flow of data Sorting and organizing data Cyber attack
Infrastructure	Modernization of outdated infrastructure and the integration of new technology
Connectivity	Simultaneously connecting a vast number of sensors Software errors or ongoing deployment issues
DT Challenges	
Data	High-quality data Noise-free data Importance of the quality and number of IoT signals
Infrastructure	A robust infrastructure is essential for the successful implementation of IoT and data analytics
Standard modeling	Ensuring a standard approach, especially for physics-based or design-based Guaranteeing information flow at every stage of DT development and implementation

2.4 Digital Twin in Power Systems: Evolution, Integration, and Data Challenges

As power grids grow increasingly complex due to factors such as the integration of renewable energy sources and decentralized power generation, the necessity for advanced digitalization becomes more apparent. Traditional SCADA systems, which have long been fundamental in monitoring and controlling power systems, are now facing limitations. While SCADA enables real-time monitoring and feedback control, it primarily functions in a reactive manner and requires manual operator intervention. Its two-way data flow is limited to control actions and does not include predictive or self-learning capabilities. In contrast, a DT represents a continuously updated and model-based digital replica of the physical power system that can simulate, forecast, and optimize its behavior. Thus, although SCADA operates “in the loop,” it cannot be regarded as a DT, since it lacks continuous synchronization, predictive modeling, and lifecycle integration. SCADA systems incorporate some level of simulation for operational purposes; however, these often lack the dynamism needed for predictive analysis and fail to support autonomous decision-making processes.

By combining Industry 4.0 methods and DT, DT in Power System (DTiPS) offers an interconnected simulation model. This model, grounded in the principles of Industry 4.0, provides a database for integrated and data-driven processes across the entire lifecycle of the power grid, from design and commissioning to operation, maintenance, and decommissioning, enabling predictive, transparent, and adaptive management of grid operations. DTiPS enhances transparency, enabling the virtual design, testing, operation, and optimization of products, processes, and systems [54]. DTiPS is a virtual representation of a real object, mapping a linkage between digital models of at least two domains. It captures attributes and functionalities throughout its lifecycle via a digital communication infrastructure, allowing for manual adjustments [55].

The Singapore Electric Utility, the pioneer in implementing the "DT for National Power Grid," incorporates two vital components in its approach: a) asset twins which manage the health of grid assets like substations, transformers, and cables, and b) network twins play a key role in assessing grid economics, security, and reliability [56].

Table 2.5 summarizes some research studies on the application of DTs in the power sector, outlining each paper's main objectives and key conclusions, ranging from enhanced grid monitoring and fault detection to innovative designs and improved operational efficiencies.

Despite the progressive digitalization of power grids, integrating real-time data across different components and systems presents substantial challenges. Currently, the power grid mirrors the industrial era characteristics known as Industry 3.0, characterized by the introduction of automation and computing. However, despite these advancements, operations remain largely siloed, with data flowing predominantly hierarchically. For instance, substations might relay

information to a central control center without significant lateral communication across similar operational levels.

Table 2.5: Overview of key research on DTs in power systems

Reference	Objective	Main conclusions from the paper
Andre Kummerow et al [57].	Monitor and control the future grid with centralized and decentralized DT	DT is used within the control center, and this technology enhances dynamic observability and anomaly detection capabilities. Reliable communication between the control center and substations requires coordinated interaction between centralized and decentralized DTs, enabling a consistent and transparent exchange of operational data.
Tao Lio et al. [58]	Create a virtual model of the power grid for grid planning and construction	By integrating DT and technologies like IoT, it is possible to enhance the planning, construction, and operation of the power grid.
Xing He et al. [21]	DTs are explained for real-time power flow analysis.	DT for power systems offers several advantages, such as real-time data integration achieved by active integration between digital and physical spaces, big data analytics in high-dimensional space, and self-adaptation enabled by data and feedback accumulation.
Mohammadi Moghadam et al. [59]	DT applications are reviewed in some parts of power, such as wind turbines, solar panels, power electronic converters, and shipboard electrical systems.	DT can be used for diagnostics, fault analysis, and control of some parts of power in real-time. For example, fault diagnosis in PV panels and a decrease in the differentiation between physical and software-in-loop controllers in wind turbines.
Baldassarre et al. [60]	A new method for the design of wind turbine blades is introduced, and a DT is developed to match the	An affordable DT model is designed for wind turbine blades, which is useful for reducing uncertainties, predicting

Reference	Objective	Main conclusions from the paper
	data obtained from the experimental test.	structural changes, and assessing remaining life.
Jain et al. [61]	A design method for fault detection is introduced by developing a DT that estimates the measurable characteristic outputs of a PV power conversion unit in real time.	The DT for fault diagnosis is able to quickly detect and precisely identify various fault types, regardless of the converter topology or type of PV installation.
Brosinsky et al. [62]	Combine Machine Learning (ML) techniques and DTs to provide new insights into security (dynamic) monitoring and control for electric power systems. It covers the following topics Combining ML and DTs for power system analysis, dynamic security assessment, probability, and risk analysis with DTs	The combination of ML and DTs has the potential to improve security monitoring and control by providing real-time decision support and accurately predicting overall system response to high-risk scenarios, while also increasing solution speed.
Yiguo Zhong. et al. [63]	Designing and implementing an innovative DT platform for a smart grid system and substation, enhancing grid management, operation, and maintenance	The innovative DT platform is used to inspect efficiency, operations, and maintenance, and it generates reports regularly, supports custom reports, and provides remote operation guidance for the monitoring center.
Michael Borth et al.[64]	Exploring DT in cyber-physical systems of systems with a focus on smart grids and smart buildings.	DT effectively addresses the unique challenges of system-of-systems engineering, providing advanced tools that enhance trust and reliability in cyber-physical systems and IoT environments.
William Danilczyk. et al[65]	Integrating DT with a real-time anomaly detection convolutional neural network and using high-fidelity measurement data.	The improved anomaly detection with DT detects the faults in the grid, and it is suitable for real-time deployment, and helps to have smart grid operations.

This hierarchical structure is shown in Figure 2.5, which illustrates the three distinct layers of the grid: the device or edge layer, the fog layer, and the cloud layer [66, 67]. At the device layer, immediate data processing allows for the monitoring of real-time power usage, voltage

levels, and equipment health, involving key components such as renewable energy resources, plug-in hybrid electric vehicles, and various sensors. The fog layer acts as an intermediary, aggregating data from various edge devices. It performs a level of intermediate processing and maintains a line of communication to the cloud layer. The cloud layer serves as the repository for historical data and executes complex analytics, overseeing the broader operational aspects of the smart power grid.

While this structure enhances efficient local decision-making and response capabilities at the edge and fog layers, it does not inherently support extensive data analysis across the wider grid. This segmentation can hinder the grid's ability to fully leverage interconnected data for more comprehensive strategic planning and operational resilience.

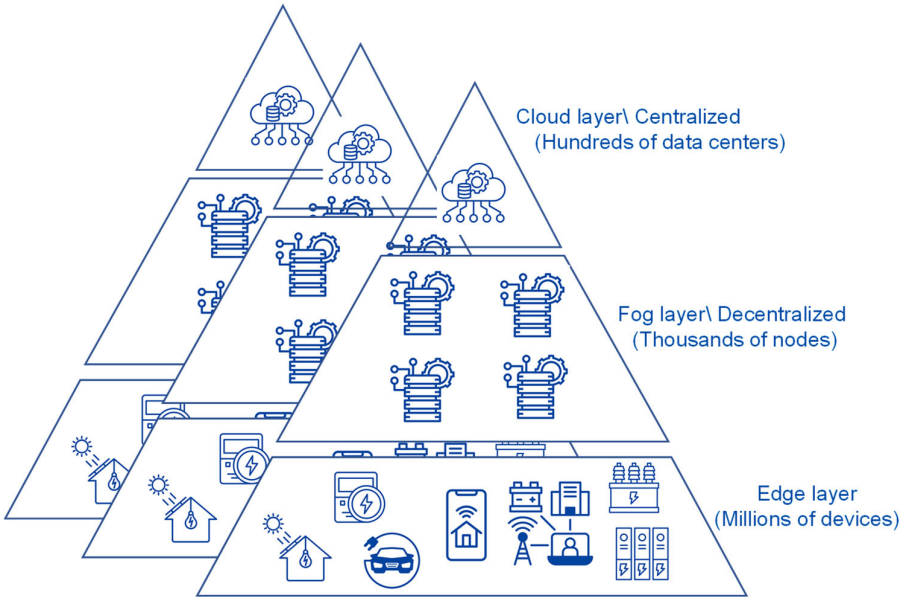


Figure 2.5: Structure of power grid (Industry 3)

DTs can revolutionize this structure by ensuring real-time integration of data from multiple sources. Figure 2.6 illustrates the evolving architecture of the modern power grid, where traditional hierarchical structures are replaced by distributed, interconnected systems. Although elements of localized intelligence already exist in conventional bay controllers and substation automation systems, DTs expand these capabilities by adding high-fidelity modeling, continuous data assimilation, and predictive analytics. As a result, functions such as planning, asset management, and operational control are no longer tied exclusively to the central control room but can be supported throughout the grid. Each component, from generation units and substations to distributed energy resources and consumers, acts as a digital node that can communicate, analyze, and respond in real time.

This architecture resembles the distributed model of Industry 4.0, where functions and intelligence are embedded across all levels of production. Similarly, in the energy sector, DTs

enable decentralized coordination, improve system flexibility, and enhance resilience by allowing autonomous decision-making close to where data is generated. Regulatory bodies such as BNetzA remain, however, outside the operational and asset-management processes; their role is limited to oversight and establishing the regulatory framework within which grid operators apply these DT-enabled functions.

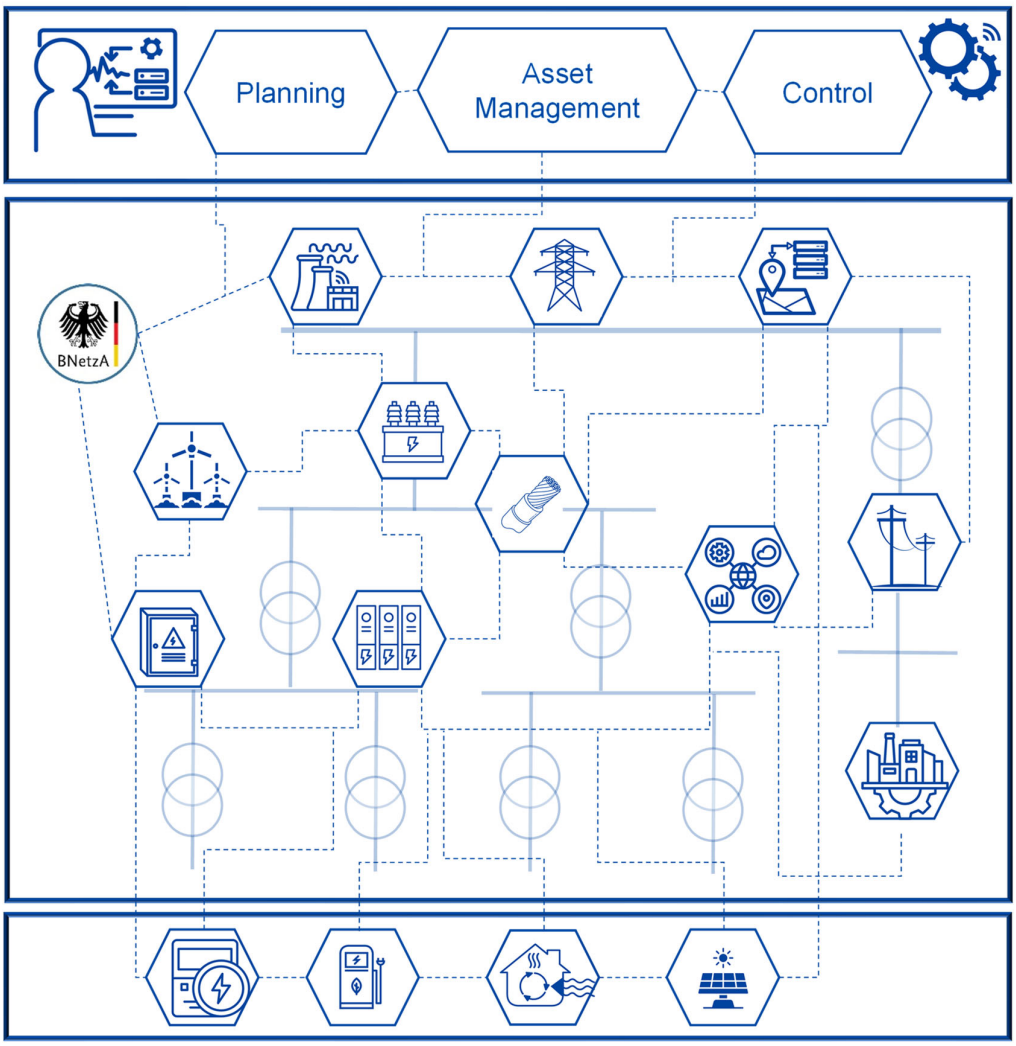


Figure 2.6: DT-enabled power grid architecture

The network infrastructure is not confined itself to a single part of the grid. Instead, it extends across grid boundaries, linking different TSOs, DSOs, suppliers, and customers. Traditionally, when updates such as changes in cables or transformers occur, it takes days for all simulation and control tools to reflect these changes. DTs can significantly accelerate this process by providing a single source of truth for grid data and enabling near-instantaneous updates across the system. Furthermore, DTs address the challenge of accessing comprehensive grid data for future planning and testing, allowing for virtual simulations that reduce costs and enhance decision-making accuracy. For example, cloud-based DT technology provides a solution for

bridging the gap in digital capabilities within the smart grid by enabling the creation of virtual models that represent existing or future physical objects [68].

Another significant issue in current grid operations is the continued reliance on predominantly time-based and reactive maintenance strategies, which only address problems after they arise. DTs facilitate a shift towards predictive maintenance by integrating real-time and historical data, thus anticipating and resolving issues before they cause failures. As customer interactions and the grid grow, managing energy consumption and production becomes increasingly complex. Today, such tasks are largely handled by grid operators and control centers, which often lack a unified, up-to-date representation of system conditions. DTs can enhance the integration of customer interactions with grid operations, improving load management, energy storage solutions, and market interactions.

The use of DTs in power system monitoring and control offers a practical solution that has the potential to advance control center technology to the next stage of evolution. Table 2.6 highlights the progress in simulation and control center technologies [69].

Table 2.6: Evolution of control center technology in power systems

	Simulation technology	**Control center technology**
1^{st} Generation	Simulation is limited to particular topics	Hard-wired, fully analog communication
2^{st} Generation	Simulation tools: A simulation is a standard tool for engineering	IP[3]/TCP[4]-based communication
3^{st} Generation	Simulation-based system design	Dynamic assessment tools
4^{st} Generation	Simulations based on interconnected data from various objects managed by DT	Data-driven support systems enabled by DT

Due to the complex nature of multiple stakeholders and subsystems in modern energy systems, it is impractical to rely on a single simulator. Instead, a holistic DT must be created by intelligently combining individual components into a comprehensive framework [70]. DTs for integrated energy systems must be robust enough to represent multi-physics systems.

These multi-domain DTs can be implemented either using a monolithic solver, which integrates all model components within a single computational framework, or by coupling several specialized models and solvers within a co-simulation environment [71].
Figure 2.7 illustrates the concept of Networked DTs in the context of the European power grid.

[3] Internet Protocol (IP)
[4] Transmission Control Protocol (TCP)

In this structure, each element of the grid, such as a power plant, substation, regional system, or operational processes such as forecasting or maintenance, is represented by its own local DT. These DTs are digitally connected, forming a coordinated, multi-layered network that mirrors the physical structure of the grid.

The figure zooms in from a continental perspective to local components, showing how DTs at different levels interact. Each DT operates independently but also communicates with others, enabling real-time data exchange, synchronized simulation, and system-wide optimization.

This interconnected DT framework allows the power grid to be managed more intelligently across all voltage levels. By working collaboratively, DTs support monitoring, forecasting, fault detection, and decision-making, both locally and across borders. Just as TSOs coordinate in the physical grid, these networked DTs enable dynamic, real-time collaboration that enhances overall grid reliability, flexibility, and resilience.

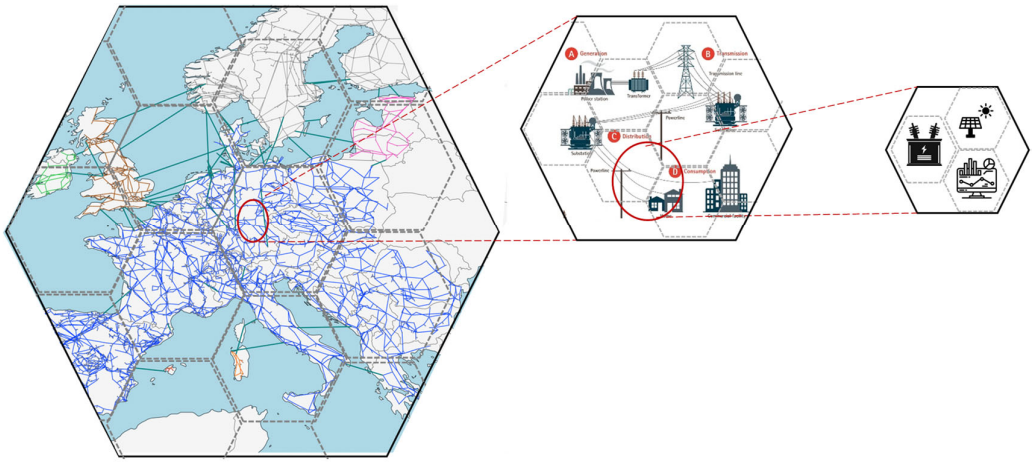


Figure 2.7: Networked DT in power grid

The structure of power networked DTs consists of multiple DTs, each responsible for modeling a specific object, system, or process. These DTs may originate from different manufacturers, making interoperability crucial for seamless data transfer. To maintain consistency and avoid discrepancies, the system requires a Single Source of Truth (SSoT) across all DTs.

DTs face significant challenges in data management, which are crucial for their effective implementation. These challenges include:

- Handling diverse data types: DTs must manage a wide range of data types that come from various stages and sources throughout the product lifecycle. The diversity of these data types can complicate integration and analysis processes.
- Managing large volumes and complexity: The large volume and complexity of the data present significant storage and processing challenges. Effective solutions must be

developed to handle these demands without compromising system performance.

- Ensuring data accuracy and integrity: Keeping data updated to accurately reflect both physical and virtual systems introduces substantial complexities. Sophisticated data management practices are required to maintain data integrity and utility [72].

DTs rely on the principle that all gathered data should originate from a single, consistent source to ensure accuracy. However, a major hurdle is the acquisition of high-quality data. Developing a robust data architecture is crucial, not only to accommodate but also to effectively manage this data in alignment with specific operational requirements to achieve desired outcomes.

In power systems, the variety and complexity of data are substantially higher compared to conventional industrial systems, as they include heterogeneous operational, planning, and maintenance data originating from numerous distributed assets. This makes managing and integrating this data particularly challenging, requiring tailored strategies to ensure seamless functionality and reliability of the DTs.

2.4.1 Data Complexity in Power Systems

With rapid advancements in digitalization and cloud computing, vast amounts of data are generated through digital devices like smartphones, computers, and sensors. The internet data volume is now measured in exabytes (10^{18}) and zettabytes (10^{21}) [73].

In the era of Industry 4.0, AI and big data integration face challenges due to data unavailability, limited focus on virtual objects, and isolated usage rather than full integration across activities. This results in disrupted data flow and poor interoperability. Key issues in industrial value chains include incompatible information for optimization and insufficient integration of product management and data exchange processes [19].

Implementing and utilizing DTs poses several challenges, many of which also occur in other digital systems such as SCADA or GIS. However, DTs typically amplify these challenges because they require higher data granularity, tighter synchronization, and continuous bidirectional data exchange. Key challenges include:

- Data acquisition (Acquiring high-quality data from sensors and meters is critical, especially in areas where meters are not readily available).
- Data storage.
- Ensuring data privacy and security.
- Protecting against cyber attacks (As digitization increases, so does the potential attack surface for cyber threats).
- Latency.
- DTs connectivity.
- Integration with existing systems.

- Workforce training to effectively use the new system.

A significant challenge for DTs is not only acquiring high-quality data but also implementing an appropriate data architecture that enables effective data management tailored to specific operational requirements and tasks. This requires substantial efforts and careful planning to ensure successful deployment and operation.

Furthermore, the need for big data approaches arises from the rapid accumulation of data and the expansion of databases. Effectively managing this vast amount of information should no longer be seen as a barrier but as an essential aspect of modern data handling. Big data is often described using the "5-V" model, which characterizes it by five key dimensions as shown in Figure 2.8 [74]:

- Volume: Describes how to handle large quantities of data, megabytes (MB, 10^6) to zettabytes (ZB, 10^{21}).
- Variety: stands for different types of data (structured, unstructured, and semi-structured).
- Velocity: Describes the speed at which data is produced, processed, and analyzed.
- Veracity: Highlights the quality and accuracy of the data.
- Value: Focuses on the informational value. (Descriptive analytics explains what has happened. Predictive analytics forecasts what is likely to happen. Prescriptive analytics suggests actions to take. Cognitive analytics uses AI techniques to learn from data and support human-like decision making.)

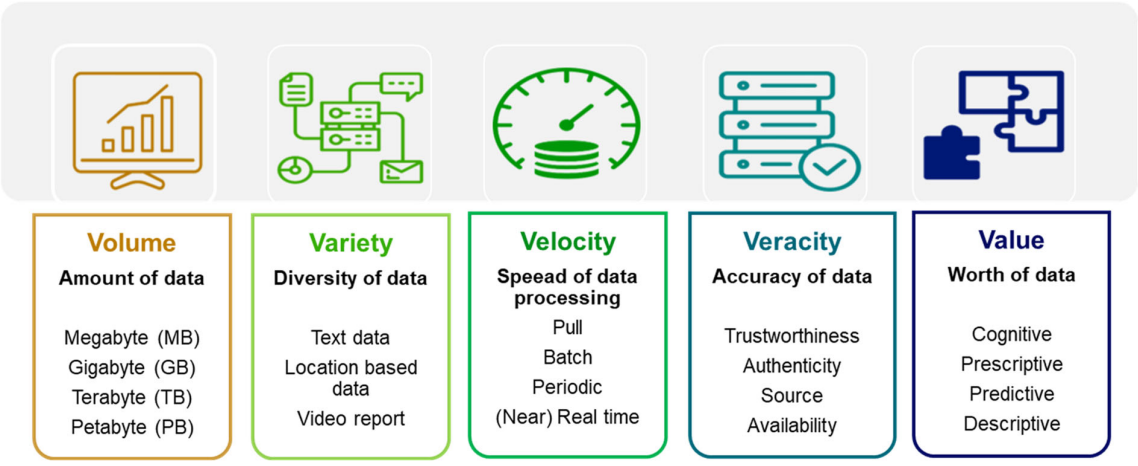


Figure 2.8: Dimensions of the 5-V model for big data

2.4.2 Big Data Characteristics in Smart Grid

Big data holds tremendous potential to enhance power grids by uncovering new opportunities and delivering diverse benefits. As power grid and communication technologies advance, they generate a vast and complex array of data. This data, collected through advanced metering

infrastructure, provides detailed electrical measurements essential for grid management and optimization. Integrating data, which spans electrical readings to weather and demographic details, is crucial for effectively managing grid operations. However, transforming these large, heterogeneous datasets into actionable insights presents significant challenges in terms of computational complexity, data security, and system integration within existing power frameworks [75].

Additionally, an average distribution utility manages thousands of terabytes of new data annually, sourced from a variety of instruments like smart meters. Other sources include remote terminal units, smart plugs, programmable thermostats, smart appliances, and sensors on critical grid-level equipment such as transformers and network switches. Data also comes from asset inventories, SCADA systems, and geographic information systems (GIS), further underscoring the complexity and scale of data management challenges in modern power grids. Table 2.7 presents the rapid growth of metering data, particularly evident when collected every 15 minutes across one million metering devices. This results in a staggering total of 35.04 billion records, with the volume of meter reading data surging to 2920 terabytes [76].

Table 2.7: The amount of data collected by 1 million metering devices in a year

Interval	**1/day**	**1/hour**	**1/30 min**	**1/15 min**
Records (billion)	0.37	8.75	17.52	35.04
Volume of data (Tb)	1.82	730	1460	2920

Data within power grids can be categorized into several types, such as measurement data, business data, static grid data, GIS data, and external data. Measurement data consists of operational parameters collected from sensors. Business data includes information on marketing strategies, competitive behavior, pricing models, and financial transactions. External data encompasses factors like weather conditions and social events, such as festivals, which impact grid operations and planning [77, 78].

Data from these various sources can also be categorized by format, as outlined in Table 2.8, structured, semi-structured, and unstructured [79]. Structured data refers to data organized in a predefined format, easy to search and analyze. Semi-structured data is not fully structured but contains some organizational markers, like JSON or XML, and unstructured data is data without a predefined format, including text, images, and videos, requiring more complex tools for processing. The growing volume of data in power grids, along with its diverse formats and sources, requires high quality and consistency. This makes the data validation and processing within DT a crucial aspect of grid management.

Table 2.8: Classification of Data Types in Power Grids

Structured data	Semi-structured data	Unstructured data
Equipment parameters Load control Meter readings GIS data	Web service data	Surveillance video footage Maintenance and inspection images

Data within power grids can be categorized into static and dynamic types, as detailed in Figure 2.9. Static data, which remains constant over time, includes information on physical assets like specifications of substations, transformers, and transmission lines. It also encompasses grid structure, represented by comprehensive maps detailing connections and layouts, and is used in simulations for statistical load flow and fault analysis to assist in planning and structural reinforcements. Dynamic data, in contrast, varies over time and is essential for real-time monitoring and operational adjustments. This category includes real-time measurements from smart meters and sensors that provide instant data on grid variables such as voltage and frequency. Additionally, dynamic data encompasses historical data that reflects past behaviors and trends, and forecasted data, along with pseudo-values that help estimate future grid conditions, supporting planning and decision-making processes.

A thorough understanding of the key measurement devices, as detailed in Table 2.9, is crucial for the effective management and analysis of the large volumes of data generated within power grids. Although these devices are essential for capturing operational information, their measurements are not perfectly accurate. The quality of the recorded data depends heavily on instrument transformers and metering equipment, which inherently introduce measurement errors due to their analog nature.

2.5 Applications of Digital Process Twin in Power Systems

DPT serves as a virtual representation of processes, enabling the identification of bottlenecks, detection of undesired events, analysis of network performance, and optimization of operations. DPTs can either replace human operators in certain tasks or assist them in areas such as data analysis, bottleneck detection, grid performance monitoring, maintenance alerts, and decision-making. In industry and machining processes, DPT is defined as a virtual representation of manufacturing processes, designed to optimize various aspects of production. Unlike DT of production systems, which focus on the overall system, DPTs aim to enhance process efficiency by ensuring optimal product quality, minimizing energy consumption, and meeting defined quality standards [80].

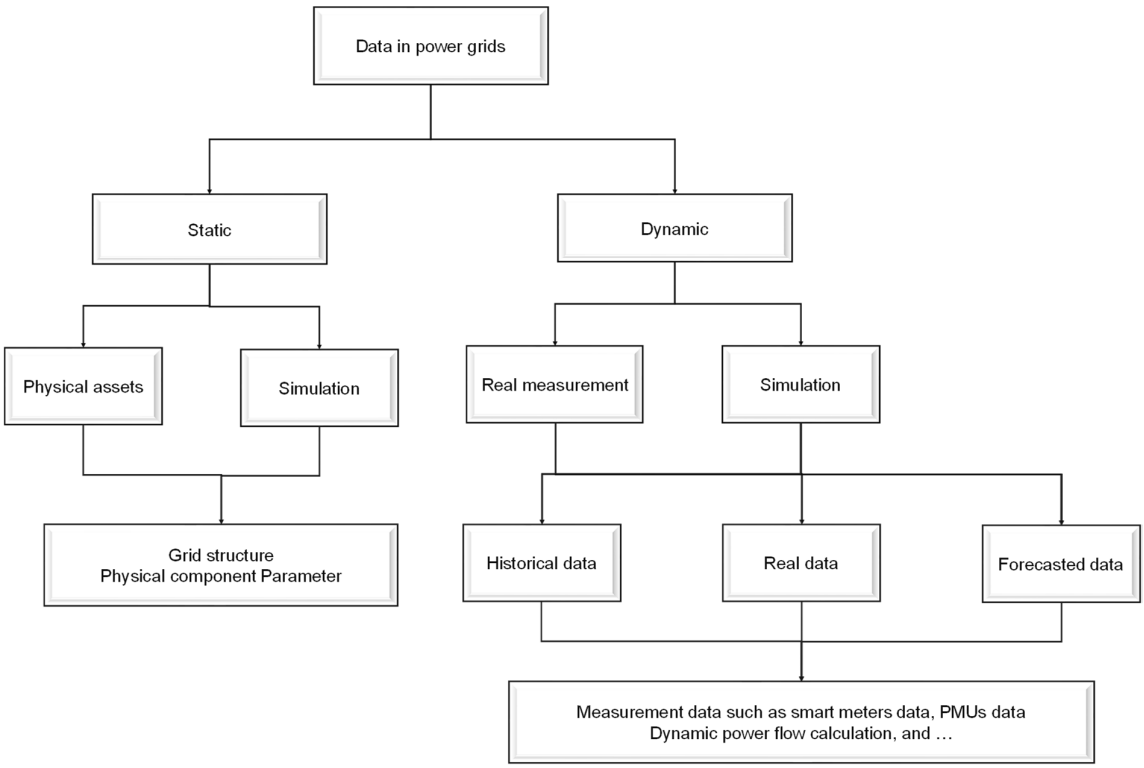


Figure 2.9: Power grid data categorization diagram

Table 2.9: Overview of key measurement devices in the power grid [81]

Sensor name	Data measured	Explanation
Power sensor	RMS[5] voltage, active power, and reactive power	The power sensor is crucial for SCADA systems and measures important electrical parameters every 2-4 seconds
Smart meter	Active power, reactive power, current, RMS voltage, power quality	Smart meters measure critical power consumption between 1 to 60 minutes
PMU[6]	3-phase voltage phasors, 3-phase current phasors, active power, reactive power, frequency, power factor	PMUs are primarily used in transmission grids but are also valuable in active distribution grids and send data at 10–60 samples per second

[5] Root Mean Square
[6] Phasor Measurement Unit

Sensor name	Data measured	Explanation
Power Quality Monitor	Frequency, RMS voltage, currents, total harmonic distortion, individual harmonics, flickers	Power quality monitors are advanced devices that capture a wide range of power quality data, 1 sample per second
Substation meters	Active power, reactive power, complex current, complex voltage	These wireless sensors are installed in transmission and distribution substations and provide data, 1 sample per minute
Wireless power line sensor	Line faults, line loads, power quality, conductor temperature	Low-cost overhead line sensors monitor distribution line status, including faults and loads. They report status every 5 minutes with significant accuracy and are easy to install on live networks.

In power systems, DPTs provide a digital model of the tasks performed by grid operators, including both operator-engineered and software-executed tasks. They facilitate the secure collection and display of grid and consumer data. Integrated within the system, DPTs are tailored to meet the specific needs of DSOs by using data from multiple sources to simulate system behavior, which enables real-time monitoring, analysis, and identification of optimization opportunities.

DPTs also promote the sharing of validated data and functions between different processes and objects, ensuring consistency and transparency in the interactions between multiple processes or entities. The data must be accurate and valid. Otherwise, it can lead to wrong decisions. To ensure the quality of the data, it must undergo a data check process, which includes unique identification and validity checks. The better the data quality, the more efficiently the system will perform.

The block diagram of the DPT in a power grid is shown in Figure 2.10 [84], and can be explained by considering its layered architecture and the various tasks it performs. The physical layer comprises smart meters, sensors, users, and external data sources, such as weather conditions, all of which impact grid performance. This layer is responsible for collecting real-time data from the grid. The gathered data are fed into the DPT system as input. The Information and Communications Technology (ICT) layer manages the communication, storage of data, and analysis of data. It is divided into several sub-layers including (a) the storage layer, which holds grid topology, operational data, and historical records, and (b) the

data verification layer, where data is validated, standardized, and checked for accuracy, it ensures that data from the physical layer is consistent and properly formatted, (c) the asset layer is where the various DTs of grid components, such as transformers, substations, and generators, are housed. These DTs simulate the behavior of each component, providing a virtual model that mirrors the grid's real-time state. This layer also includes simulation tools, and (d) the knowledge extraction layer is crucial for analyzing the data and simulations produced by the previous layers. Advanced analytics are carried out here, such as data analysis, load flow forecasting, risk state identification, and anomaly detection. The goal of this layer is to extract actionable insights that improve the grid's visibility and observability. These insights are crucial for identifying opportunities for optimization, detecting potential issues, and forecasting future conditions. The outputs of the DPT, such as insights, signals, and alerts, are sent to the human interaction layer, where operators and various grid management use cases come into play. This layer involves several key functions, such as grid control, maintenance management, and grid management. Others can write what they feel necessary.

2.6 Enhancing Power System Performance with Digital Twin and Digital Process Twin

DTs and DPTs offer a wide range of advantages, depending on when and where they are used. These benefits span several aspects:

- Reducing costs at various levels of the system, such as in operations, maintenance, and control.
- Enhancing system monitoring capabilities for improved visibility and control.
- Increasing the reliability and availability of the system by minimizing the risk of errors or failures.
- Improving performance through advanced analytics and data-driven insights.

The interconnected nature of multiple DTs enhances the overall performance and efficiency of power systems, offering a comprehensive view of the entire network. DT and DPT in power systems are designed to achieve multiple goals, enhancing not only the system's operational capacity but also its strategic planning [82–85]:

- Enhance monitoring capabilities to ensure continuous operational visibility.
- Increase SA through real-time data integration.
- Optimize grid performance for efficiency and resilience.
- Facilitate virtual testing for better operational planning decisions.
- Manage emergencies and urgent system conditions more effectively.
- Improve maintenance protocols and asset management strategies.
- Enhance re-parametrization and control systems.

- Reduce the risk of power outages and improve fault recovery times.
- Proactively identify and address potential system vulnerabilities.
- Achieve significant time savings in grid model creation and maintenance, with over 90% time reduction reported in some cases (reported by SIEMENS).

Synchronization of a DT with its physical counterpart is necessary to maintain high levels of SA. This is achieved by integrating predictive models and physical and analytical data, and increasing decision-making capabilities. Such an advanced application of DT not only supports decisions but can also automate them, leading to more prescriptive solutions and proactive management of the power grid. This ensures that decision-making is both timely and data-driven, improving the outcome based on the comprehensive SA provided by DT [86].

2.7 Summary and Research Focus

DT and DPT technologies offer numerous advantages for modern power systems, including improved monitoring, control, and decision-making. While DPTs cover a wide range of functions and components, this thesis focuses specifically on the knowledge extraction layer, which processes real-time data to generate operational signals. These signals are crucial for enhancing SA in LV grids. The next chapter presents the sublayers and methods within this layer, anomaly detection and forecasting, that are developed and applied to improve SA.

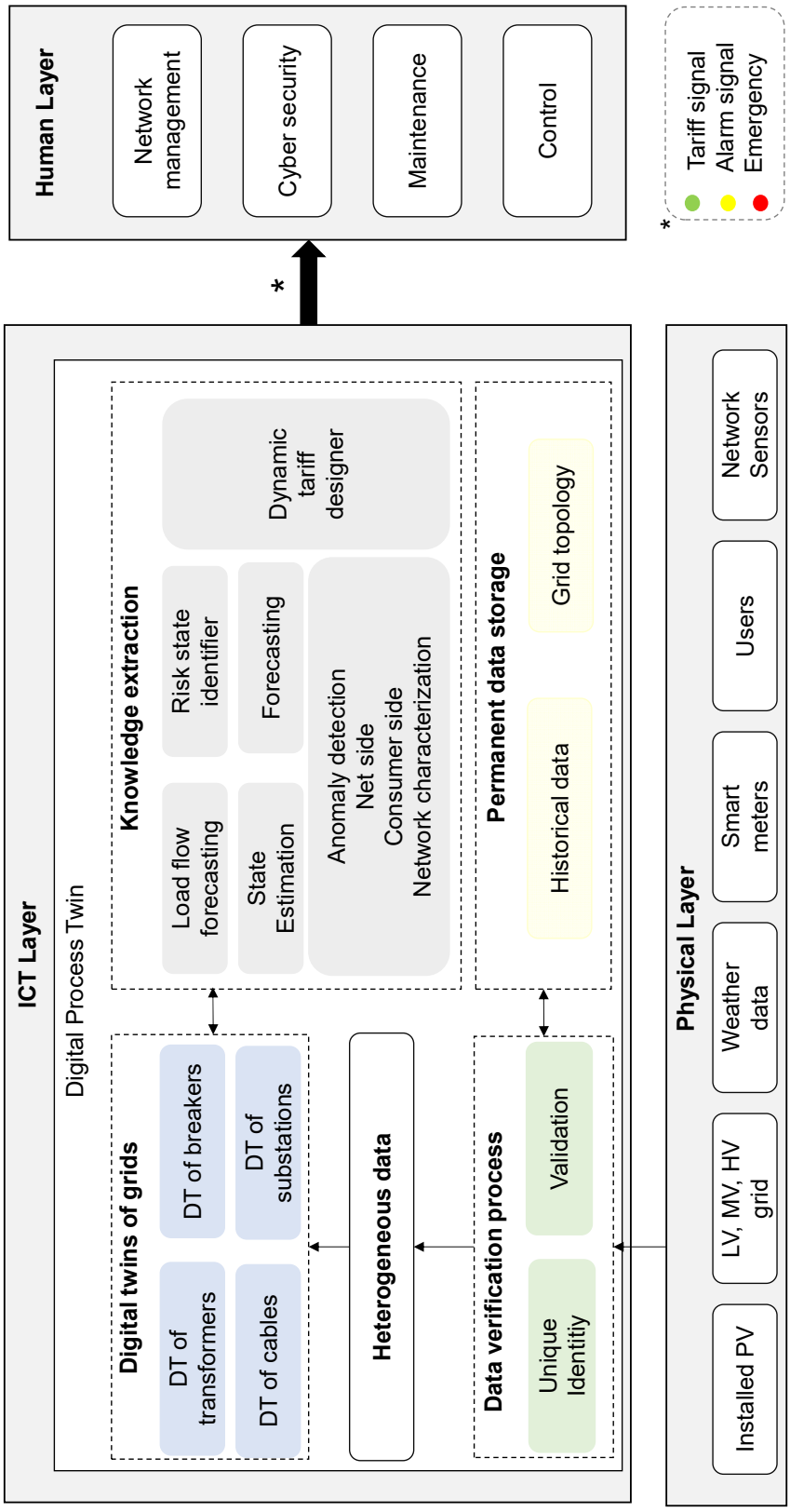


Figure 2.10: DPT structure in power grid [84]

3 Situational Awareness in Power Systems

Chapter 2 introduced the concept of the digital process twin, where knowledge extraction plays a key role, particularly in improving Situational Awareness (SA). This chapter focuses on defining SA and highlighting its importance for power system reliability, especially in distribution grids. In contrast to medium-voltage and transmission systems, LV grids operate much closer to customers and therefore show more diverse and rapidly changing demand patterns, but they have far fewer sensors and much less real-time information available. These characteristics make achieving SA in LV networks significantly more challenging. Due to measurement limitations, model inaccuracies, and a high frequency of events, SA in LV grids faces unique challenges. Various methods have been proposed to enhance SA, and in this thesis, the focus is placed on preventive forecasting and anomaly detection at the customer level to support early decision-making and improve system resilience.

3.1 Situational Awareness: A Key to Power System Reliability

SA was first introduced by Endsley in 1988 [87] and is defined as the ability to perceive environmental elements over some time and space, understand their significance, and anticipate their future status. Initial studies on SA primarily occurred within the aviation and military sectors [88]. Real-time SA stems from exercises in the US Air Force designed to train fighter pilots to anticipate enemy maneuvers [89]. This approach involves continuously collecting data about the current situation and environment, interpreting what this data shows, and predicting future conditions to formulate appropriate actions. This process is enclosed by the military observation-oriented decision-action loop. In the context of power systems, SA is defined as the ability of system operators to perceive, comprehend, and predict the elements within the environment over time and space. This capability is crucial for effectively identifying anomalies and responding to dynamic operational challenges [90].

The lack of SA was critically exposed during the August 14, 2003, blackout in the United States and Canada, which was attributed to four main causes:

(1) Inadequate system understanding.

(2) Inadequate SA.

(3) Inadequate tree trimming.

(4) Inadequate support from reliability coordinators for diagnostics.

The collapse in SA was starkly evident when First Energy's monitoring systems failed, subsequently disabling the control room consoles from receiving any further alarms. This lack of audible and on-screen alerts prevented operators from recognizing significant changes in the system’s conditions, a failure that was critical during the crisis. It was suggested that stricter reliability standards be implemented to address these vulnerabilities and strengthen the

network's physical and cyber security. The recommendations emphasized the critical need for improved SA, not only through the monitoring and quick response to initial signs of system distress but also by managing responses through well-trained personnel and robust diagnostic support systems. The enhancement of monitoring tools, crucial for detecting issues such as meter failures and promptly providing assist signals, was highlighted as essential to prevent similar incidents in the future, ensuring grid stability and reliability through effective detection and response mechanisms [91].

Blackouts are generally clustered in two phases: (a) the pre-cascade phase, where operators can react and implement control measures, and (b) the cascade phase, characterized by rapid component failures that are beyond the operator's ability to respond. In the critical pre-cascade phase, maintaining control is critical and requires strong SA and the availability of advanced decision support and monitoring tools. These tools help operators quickly collect and analyze data to effectively manage a situation before it gets out of hand [92].

According to reports by the NERC[7] Disturbance Analysis Working Group, aside from weather-related incidents, operator errors are the second most prevalent cause of disturbances, closely following equipment failures [93]. By enhancing SA, information systems can better support human decision-making, providing operators with critical assist alarms that help prevent errors before they lead to system disturbances. SA is widely defined as the perception of environmental elements within a specific timeframe and space, understanding their significance, and predicting their future status, as illustrated in Figure 3.1 [94]. The output generated at the projection stage is a signal or alarm that assists in decision-making. Based on this output, appropriate actions are taken, and feedback is subsequently sent back to the system for ongoing refinement.

- Perception: This initial stage emphasizes real-time monitoring and data collection, which are essential for understanding the current state of the grid. Key activities include data acquisition, environmental sensing, and information prioritization to ensure that the most relevant data is highlighted for further analysis.
- Comprehension: At this stage, the focus shifts to processing and analyzing the acquired data to gain a deeper understanding of the current grid conditions. This involves synthesizing data through analysis, integrating it within the existing context, and recognizing patterns that may indicate underlying issues or opportunities.
- Projection: This level uses predictive models to forecast future grid conditions based on current data and identified trends. It also involves the detection of potential issues before they materialize, allowing for proactive measures to mitigate risks effectively.

SA research in power systems is still developing; it plays a vital role in enhancing the situational

[7] North American Electric Reliability Corporation (NERC), https://www.nerc.com

capabilities of control centers. By improving perception and data integration, SA tools can significantly strengthen grid reliability and reduce operational risks [95].

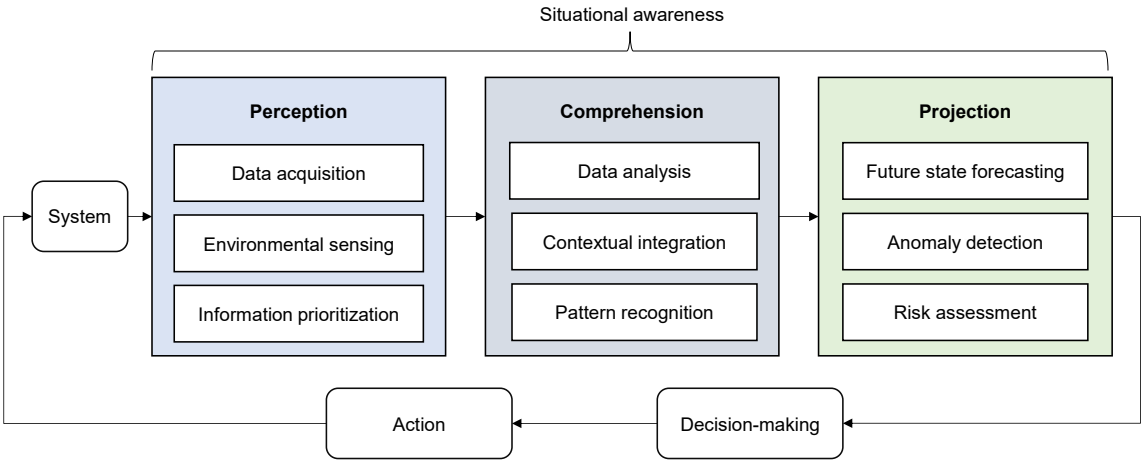


Figure 3.1: Structure and functions of situational awareness

These capabilities not only support day-to-day operation but also improve the grid's resilience during natural disasters. In summary, SA serves as a cornerstone for modern power system operation, enabling operators to perceive, interpret, and proactively respond to changing conditions. Its absence can lead to critical failures, while its enhancement supports both resilience and reliability in an increasingly complex grid environment.

3.2 Methodology to Improve Situational Awareness

Achieving SA in power distribution systems faces significant challenges [96], such as:

a) Measurement limitations: Most data from distribution systems is collected through SCADA systems at substations, which update on a minute-by-minute basis. However, smart meters, which provide more frequent residential data, typically only report every 15 minutes or even hourly, reducing the granularity of the data available. This lack of high-resolution data restricts the accuracy of real-time monitoring and event detection.

b) Model inaccuracies: A major issue is the lack of current and accurate models for most distribution circuits. This deficiency hinders effective system management and accurate predictive analysis, which are essential for operational efficiency and proactive maintenance.

c) High frequency of system events: The lower voltage levels and the greater diversity in utility and customer equipment within distribution grids lead to a significantly higher number of daily events. This volume of activity adds complexity to system management and operational reliability.

To enhance SA within power distribution systems, several solutions have been proposed, each targeting specific challenges: The first challenge, measurement limitations, has been partially mitigated by deploying more meters across the grid [97]. Although these technologies are not

yet widely implemented across all distribution networks, these units provide high-frequency data, significantly improving the data's resolution. The issue of model inaccuracies can be addressed through advanced algorithms that detect and adapt to changes in grid topology [98] or using methods that are independent of the grid topology (model-free), such as data-driven and artificial intelligence-based approaches. Further progress is expected in the integration of DT into the grids, which encompasses all grid components and is continuously updated as changes occur. This approach ensures that the models used are as accurate and up-to-date as possible.

3.2.1 Challenges and Techniques in Forecasting for Low-Voltage Grids

In LV grids, the rise of decentralized energy sources and new loads introduces variability and uncertainty. Forecasting for LV grids presents significant challenges primarily due to the nature of the data involved. The diversity of data sources introduces high variability and uncertainty in demand patterns. Many LV grids also suffer from a lack of comprehensive historical data, which is essential for developing accurate forecasting models. The intermittent nature of renewable energy sources, like solar and wind, adds complexity to forecasting efforts, as these sources are heavily influenced by changing weather conditions. These complexities necessitate the use of advanced probabilistic methods to manage unpredictability effectively. The forecasting of LV grids serves several critical purposes:

- Operational efficiency: Accurate load forecasting enhances the operational efficiency of power systems, ensuring that electricity generation matches demand. This is essential for keeping the system stable and avoiding power outages [99]. It provides additional insight into local consumption patterns that can support more refined operational strategies in distribution networks.
- Decision-making: Forecasting aids in informed decision-making for optimizing power systems. It helps operators prepare for and manage changes in load, which is crucial for both immediate and future planning [100]. Although not all decisions in LV grids require explicit forecasting, short-term predictions of local loads can enhance planning for congestion management, transformer loading, and flexibility allocation.
- Integration of renewable energy sources: As renewable energy sources become increasingly prevalent, accurate forecasting becomes essential for managing the inherent variability and uncertainty they introduce. This ensures the grid remains reliable and that the energy supply can meet demand even as conditions change [101].
- Voltage control: Managing and mitigating issues like over-voltages and voltage drops are crucial in grids with high penetration of photovoltaic systems.
- Congestion management: Forecasting supports the early identification of potential grid congestion by predicting intra-hour load and generation patterns [102].

Forecasting in LV grids can be categorized along two dimensions: the time horizon and the level of spatial granularity. The time horizon categorizes forecasts into four subcategories: very short-term, short-term, medium-term, and long-term, each reflecting the length of time into the future being forecasted, from minutes to months [103, 104]. The spatial granularity refers to the aggregation level of the load data, which can range from individual households to entire regions [105]. Moreover, forecasting methods are divided into three main types: parametric, nonparametric, and intelligence-based methods [106]. Parametric methods use mathematical models to establish relationships between variables such as power demand, temperature, and natural illumination [107]. Nonparametric methods rely on historical data to identify patterns, such as repetitive demand patterns influenced by temperature and days of the week [108], or use techniques like kernel density estimation [109]. Intelligence-based methods employ neural networks, deep learning, and other artificial intelligence technologies to handle complex and variable data, such as smoothed voltage profiles or aggregated data from multiple nodes [110, 111]. This approach is particularly useful where traditional models fail to capture the high fluctuations in voltage profiles at individual nodes.

3.2.2 Anomaly Detection in Power Systems

Anomaly detection is a critical process that identifies unusual patterns in the observed data that deviate from the expected behavior. These atypical patterns are known as anomalies, outliers, or exceptions. There is no absolute definition of anomaly, but there are some classic definitions of anomaly as follows [112]:

- “Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior” [113].
- “An anomaly is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [114].

There are different types of anomalies in data analysis: (a) point anomalies, which are single data points that significantly deviate from the rest, (b) contextual anomalies, which depend on the context of occurrence, for example, the energy data might be normal during the day and could be abnormal during the night, and (c) collective anomalies, which involve a group of data points that are classified as anomalies in the entire dataset. Outliers, though similar to anomalies, are defined slightly differently; they are data points that significantly deviate from other observations in a dataset. In a time series, outliers may not impact the overall pattern or trend of the data. Figure 3.2 shows all types of anomalies and outliers. Anomaly detection in power systems is an essential area of research that addresses operational and security challenges by identifying anomalies manifesting as fault detection [115], Cybersecurity attack detection [116], and energy theft detection [117, 118]. Deviations from baseline bidding behaviors in smart grids are explored to indicate potential cyber threats [119]. Anomalies can occur at different levels of the power system and involve various types of data. For instance,

anomalies may appear in voltage values, frequency fluctuations, or unusual patterns in energy consumption, which can occur at different levels of the system and affect a wide range of data points, from voltage fluctuations to unusual patterns in energy consumption. One study, referenced as [120], delves into individual household behaviors to identify abnormal energy consumption in buildings. Other research, as referenced in [121], targets anomalies in PV power generation, aiming to pinpoint anomalies directly related to solar energy output.

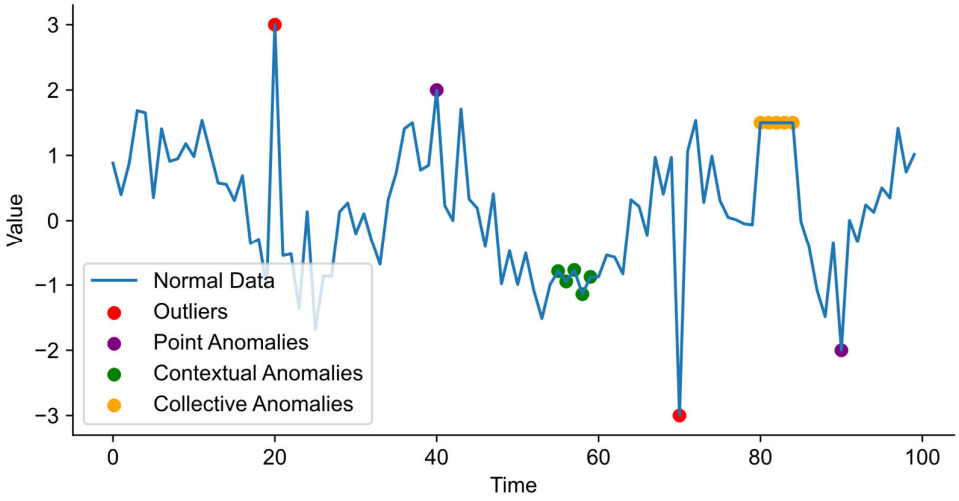


Figure 3.2: Time series with different types of anomalies and outliers

Various advanced Machine Learning (ML) techniques have been employed in this area, demonstrating the range of applications and effectiveness. For example, autoencoder schemes have been tailored for anomaly detection in energy management to preempt anomaly states using hourly power consumption data [122]. Bi-directional Long Short-Term Memory (LSTM) autoencoders analyze metering data from energy sources across 985 households [123]. Additionally, LSTM autoencoders have been deployed for cybersecurity attack detection in smart grids, involving the IEEE 14 and 118-bus systems [124], and an LSTM recurrent neural network-based autoencoder for DC fault detection in naval shipboard power systems [125]. A convolutional neural network-LSTM-based autoencoder is used for detecting false data injection attacks [126].

Defining what constitutes normal behavior in data can be challenging, particularly in LV grids. Different anomalies could be considered in power grids, which can be categorized into three groups, as shown in Figure 3.3:

- Equipment anomalies: Equipment anomalies occur when the physical equipment or infrastructure in the grid experiences a fault or deviates from its normal operating conditions. For example, circuit breaker faults, corrosion, or wear on lines or equipment. The predominant type of equipment anomaly is the fault event. In distribution systems, fault events can be classified into four primary types: single line-to-ground fault is the most common type of fault, occurring when one of the three-phase

conductors contacts the ground, often due to environmental factors like wind, animal contact, or a falling conductor. Statistically, single line-to-ground faults represent about 70% of all faults. Line-to-line faults and double line-to-ground faults occur less frequently, comprising about 15% and 10% of faults respectively, often caused by high winds or other phase contacts. Three-phase to ground fault is the least common, occurring in about 5% of cases, typically due to equipment failures or structural collapses.[127].

- Operational anomalies: Operational anomalies refer to unexpected deviations in the functioning of the grid, often related to power flow, grid stability, or control systems. For example, voltage fluctuations due to load imbalances or equipment failures, frequency deviations, phase imbalance, communication failures, cybersecurity attacks, congestion, and islanding. Operational anomalies also encompass power quality issues, such as voltage sags, swells, interruptions, and harmonics, which can be induced by faults, large equipment operations, or devices like inverters. Managing these disturbances is crucial for maintaining a reliable power supply [128, 129].
- Consumer-side anomalies: Consumer-side anomalies arise from irregular power consumption and/or generation or problems with end-user equipment that can affect the grid and consumer devices. One growing source of these anomalies is the increasing presence of distributed renewable energy sources, such as rooftop PV systems installed in households. For example, anomalies can be caused by factors such as unregistered PV systems, changes in consumption patterns, or meter failures. Several factors contribute to these anomalies, including shifts in weather, which impact both consumption and generation patterns, fluctuations in electricity prices, and major events like the COVID-19 pandemic or large public gatherings such as football matches. Additionally, the increased use of new technologies, such as electric vehicles and electric heat pumps, can significantly alter the demand on the grid. While the long-term adoption of technologies such as electric vehicles and heat pumps represents a structural change rather than an anomaly, their initial impact on consumption can temporarily appear as an anomaly when the new demand pattern first emerges. Once the new behavioural pattern is observed, learned, and incorporated into the forecasting model, it should no longer be treated as an anomaly but as part of the updated “normal” load profile.

Anomaly detection within LV grids employs various methods, each tailored to specific types of data and anomalies. Techniques such as graph matching, useful for detecting anomalies in electric topological and configuration databases, illustrate traditional applications of anomaly detection methods as referenced in [130]. Meanwhile, real-time anomaly detection leverages models like LSTM to analyze voltage magnitude measurements, showcasing their practical utility in [131]. These methods are broadly categorized into two main types: traditional statistic

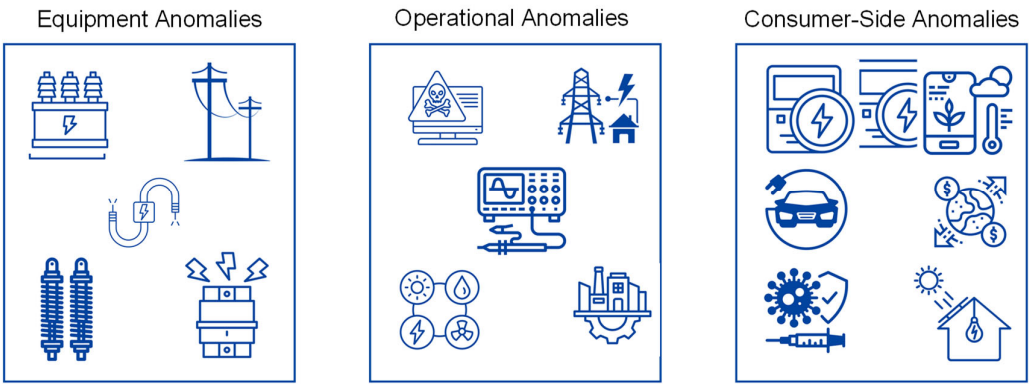


Figure 3.3: Different types of anomalies in power grids.

methods and AI-based methods. Traditional statistical methods are the oldest algorithms for detecting anomalies [132] and involve straightforward techniques that check for significant deviations from established norms. These methods are quick and simple to implement but may not effectively handle rare or complex anomaly patterns. On the other hand, AI-based methods, including ML and deep learning, are gaining popularity due to their advanced capability to learn and adapt from data. Methods for anomaly detection are illustrated in Figure 3.4.

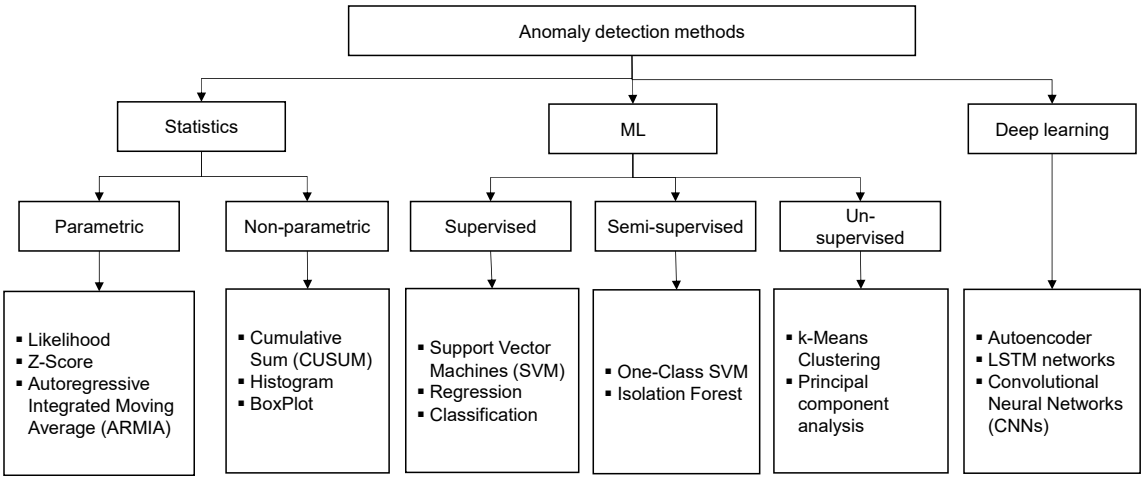


Figure 3.4: Methods for anomaly detection

ML methods excel where data complexities require robust analysis, learning from examples to improve anomaly detection. However, it is crucial to note that ML can generate errors if not properly trained. Deep learning methods go a step further by efficiently managing data with intricate or evolving patterns, often uncovering hidden structures that elude other techniques. Despite their strengths, deep learning models demand substantial computational resources and can present challenges in interpretability. In the context of ML-based anomaly detection, the approaches can be differentiated based on the type of learning involved:

- Supervised learning: This method relies on labeled data, where each record is annotated as normal or anomalous.
- Semi-supervised learning: This approach uses a mix of labeled and unlabeled data. Models are first trained on normal behavior and then classify new data as normal or anomalous based on deviations from this norm.
- Unsupervised learning: In unsupervised learning, no labels are provided, and the model is trained on the entire dataset. This method is particularly useful in scenarios where anomalies are not known a priori or are too rare to be labeled effectively.

Selecting the appropriate method for anomaly detection in LV grids is crucial due to the diverse nature of anomalies, ranging from abrupt disturbances caused by failures to gradual shifts in consumption patterns influenced by new loads such as electric vehicles and electric heat pumps. To effectively address these variations, different ML models are employed, each suited to specific types of data and anomalies. For instance, decision trees and random forests are effective for varied data types and excel in classification tasks. Deep learning models are adept at capturing complex, non-linear relationships within large datasets, while LSTM networks specialize in identifying anomalies in sequential data. Clustering algorithms like K-Means are utilized to identify deviations from common behavioral patterns, indicating potential anomalies.

The effectiveness of a chosen method depends heavily on the dataset's characteristics, such as its distribution, dimensionality, and the level of noise. Methods vary in their precision and recall; some may detect a higher rate of true positives, whereas others are better at capturing as many anomalies as possible. Isolation Forest, for example, is particularly effective in noisy environments and capable of handling missing data, making it suitable for real-world applications where data quality may be compromised. On the other hand, LSTM networks can adjust to new and evolving patterns due to their ability to remember long-term dependencies, proving invaluable in dynamic settings.

Moreover, the application of ML methods in anomaly detection is heavily influenced by the quality of the data and the specific objectives of the analysis. The unpredictable nature of power consumption and PV generation in LV grids means that training data might include unnoticed small-scale anomalies. Real data often contains uncertainties and may not always be sampled consistently. Additionally, incorporating external data such as weather conditions introduces further complexity, particularly when critical variables like global radiation are unavailable or incomplete. These challenges necessitate flexible ML methods that can adapt to continually changing grid data and consumption patterns, ensuring the detection system remains effective in a complex and evolving energy landscape.

The utilization of real LV grid data for developing AI models presents several significant challenges that can impact the accuracy and reliability of anomaly detection:

1. Variability in data: The LV grid data exhibits high levels of fluctuation and inconsistent patterns over the years, complicating the ability of ML models to discern clear trends or establish stable learning patterns. This variability often leads to potential inaccuracies in model predictions.
2. Anomalies and noise: The presence of diverse anomalies within the dataset, ranging from clear outliers to subtle discrepancies, poses difficulties. While more apparent anomalies can be addressed during preprocessing, smaller, less conspicuous anomalies often remain undetected in the training data, contributing to noisy and less accurate modeling outcomes.
3. Black-box nature of ML models: Many advanced ML models, particularly those based on deep learning, operate as "black boxes" with limited transparency, complicating the interpretation of why specific anomalies are detected. This challenge is exacerbated when models are trained on data replete with noise and subtle anomalies, leading to detections that are difficult to justify or explain.
4. External data: Often, anomaly detection is augmented by external data such as weather conditions, which may not always be accurate or fully representative of the local variables affecting the LV grid. For example, weather data sourced from distant stations may not accurately reflect local phenomena crucial for applications like PV generation, thereby hindering precise anomaly analysis.
5. Sensitivity variations: ML models differ in their sensitivity to anomalies, resulting in discrepancies in detection across different systems. Some models might identify certain data points as anomalies, while others overlook them. Although some anomalies can be rationalized by factors such as weather conditions or historical trends, others remain ambiguous due to the inherent noise and errors in the data.

3.2.3 Summary and Research Focus

This chapter explored the role of SA in enhancing the reliability of LV grids, focusing on challenges such as limited measurement resolution, inaccurate grid models, and the high frequency of operational events. Among various approaches to improve SA, this work concentrates on data-driven methods, specifically preventive forecasting and consumer-side anomaly detection. These methods represent only a subset of the broader SA enhancement strategies, but are particularly valuable due to their ability to operate independently of the underlying grid topology. By leveraging real-time and historical data, they contribute to improved awareness without relying on complete or up-to-date grid models, making them especially suitable for practical applications within the DPT framework.

4 Methodology for Enhancing Situational Awareness in LV Grids

This chapter outlines the methodology used to enhance Situational Awareness (SA) in low-voltage (LV) grids through data-driven forecasting and anomaly detection, integrated within the Digital Process Twin (DPT) framework. The methods are designed to operate independently of grid topology and rely on historical and real-time measurement data. The focus lies on two core components: Pseudo-Worst-Case Forecasting (PWCF) for preventive control, and anomaly detection to identify irregularities on the consumer side. The chapter covers data preprocessing, Neural Network (NN)–based forecasting, Machine Learning (ML), and statistical detection techniques, as well as specific applications such as unregistered PV and outage detection.

Figure 4.1 provides a structured overview of the methods employed in this thesis to enhance SA. It distinguishes between forecasting and anomaly detection tasks, categorizing each technique according to its underlying approach: AI-based, statistical, or hybrid.

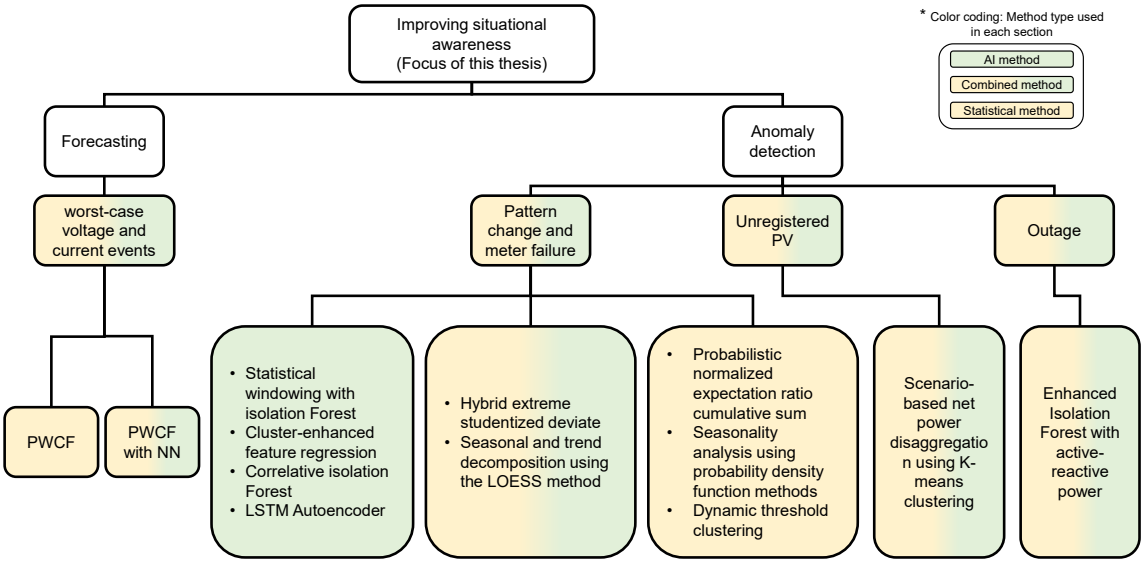


Figure 4.1: An overview of the methodological framework used to enhance SA in this thesis

4.1 Preprocessing and Analyzing Data

Data preprocessing is a critical step in data analysis, involving data cleaning, transforming raw data into a usable format, and reducing dimensionality. This process ensures accuracy, consistency, and relevance, essential for effective data analysis. Key activities in data preprocessing include:

- Data cleaning: Correcting errors like missing values or duplicates.
- Data transformation: Modifying data into an analyzable format, such as normalizing or encoding variables.
- Data reduction: Choosing a subset of data for analysis, which is crucial for tasks like

feature selection or dimensionality reduction. In high-dimensional data analysis, big data enables the discovery of hidden structures in subpopulations and the extraction of common features across diverse groups, despite significant variation [133, 134].

In industries such as power systems, effective data preprocessing ensures the accuracy, consistency, and reliability of analysis results.

Given the challenges with real data in power systems, such as sparse or incomplete ground truth, it is often difficult to verify whether the measurements accurately reflect actual system failures. For example, rising smart meter failure rates, from 1% to 3.7%, can lead to false alarms and operational difficulties [135], and the complexities of managing vast, fast, and diverse data streams from smart grids, effective preprocessing and analysis are essential. These data issues, including frequent missing values that compromise data quality and affect the trustworthiness of analyses, underscore the need for sophisticated data management strategies to ensure accurate predictions and prevent financial penalties for suppliers [136].

For this thesis, all available data have been preprocessed and analyzed, ensuring that important features and additional relevant information, such as weather information, are selected based on the specific objectives of the method. Infrequent missing values were replaced using pseudo-value generation, while long-term gaps, such as extended smart meter outages, were removed from the dataset. Data formatting was adjusted as required to match the needs of the analysis method.

4.2 Pseudo-Worst-Case Forecasting Method

The rapid integration of decentralized renewable energy sources, electric vehicles, and advanced heating systems into LV grids has significantly increased the complexity of grid management. While environmentally beneficial, these advancements present several challenges, including unpredictable voltage fluctuations, thermal stress on infrastructure, and potential overloads. Effective monitoring and control of LV grids are crucial to mitigating these issues. However, the current infrastructure, particularly the advanced metering systems such as smart meters, is not fully equipped to handle real-time data acquisition, which is essential for making timely and accurate control decisions. The data provided by these systems typically have long sampling intervals, often 15 minutes or more [137], and consist only of snapshot values, which offer limited insight into the dynamic state of the grid. Given these limitations, traditional forecasting methods that attempt to predict precise future values have proven inadequate, especially in the context of LV grids, where the stochastic behavior of prosumers introduces significant volatility. This chapter presents the PWCF method [138, 139] to solve these challenges. The method proposed in this chapter introduces a novel heuristic approach to address forecasting challenges. This method does not aim to predict exact future values. Instead, it forecasts a range within which the grid parameters are likely to fall, thus providing a practical approach to enhancing the monitoring and control capabilities of LV grids.

Forecasting voltage and current in LV grids is critical for maintaining secure and reliable grid operation and preventing potential failures. However, achieving precise forecasts in this context is particularly challenging due to several reasons, such as:

- High fluctuation and unpredictability: The behavior of prosumers, entities that both consume and produce electricity, leads to significant fluctuations in voltage and current levels. These fluctuations make it nearly impossible to predict the exact values of voltage and current at each node in the LV grid, especially in short-term forecasting.
- Impact of MV grid control commands: The dynamic interaction between LV grids and upstream MV grids adds another layer of complexity. Control commands from the MV grid can cause immediate and unpredictable changes in the LV grid, making it even more challenging to forecast exact values.
- Limitations of smart meters: Smart meters provide data at relatively long sampling intervals and often lack the detail required for accurate forecasting.

Despite these challenges, forecasting remains essential for identifying and preventing potential issues, such as voltage violations or current overloads, that could compromise grid reliability. Unlike transmission systems, control methods in LV grids typically try to prevent undesired system conditions only, which can be forecasted by the PWCF even if the nodes show highly fluctuating voltage profiles. The goal is to maintain grid parameters within secure operational boundaries, specifically, within a ±10% deviation from the nominal voltage [140] and below the maximum allowable current threshold, which is the permissible current for each line. For PWCF, data are divided into three zones for voltage values and two zones for current values based on the higher and lower permissible voltage borders (10% plus and minus the nominal voltage). Higher than the upper limit, inside the limit, and below the lower limit. The zone between limits is defined as a “non-critical zone”. While short-term volatility and measurement uncertainties exist, operation within this zone generally remains within secure operational boundaries. Therefore, in the context of preventive control, precise forecasting is less critical compared to boundary regions near the permissible limits. Figure 4.2 illustrates the segmentation of voltage values into these three zones, emphasizing the “non-critical zone” and its boundaries.

The proposed method in this chapter of the thesis tackles forecasting challenges not by aiming to make a forecast as precise as possible, but by providing a range in which the true value could lie. The term “pseudo” reflects the reality that the absolute worst-case scenario cannot be predicted with certainty due to the inherent unpredictability of the grid. The term “worst-case” refers to forecasting the upper and lower bounds of voltage and current levels.

Figure 4.3 illustrates the structure of the PWCF method. At its core, the method calculates a basic pattern ($\boldsymbol{v}^{i}_{base}$) based on the mean values of historical data, which represent the expected behavior of the consumers under normal conditions. To account for the stochastic

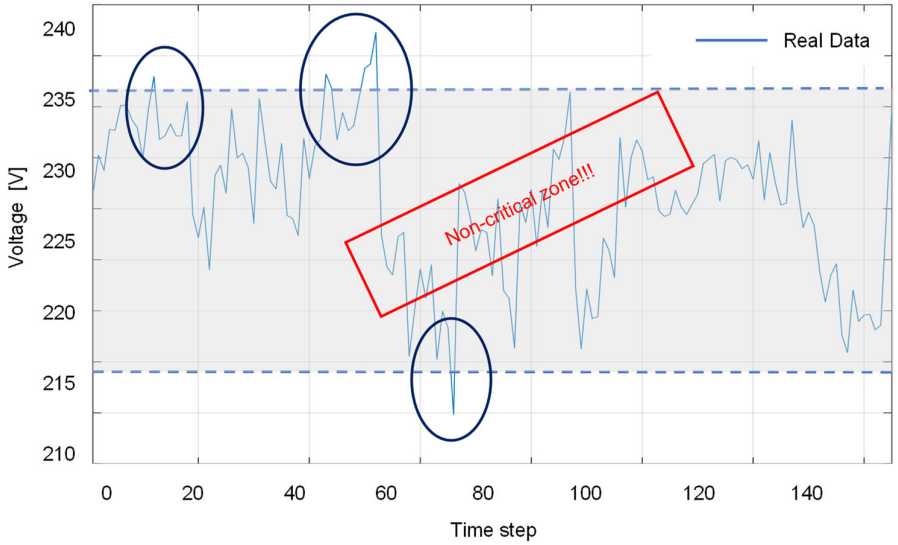


Figure 4.2: Voltage classification based on permissible limits

nature of LV grid dynamics, this basic pattern is symmetrically shifted upward ($\boldsymbol{v}^i_{u,border}$) and downward ($\boldsymbol{v}^i_{l,border}$), forming a bounded range. These boundaries serve as forecasts for the next sampling interval and are continuously adjusted to account for dynamic system behavior. The adjusted values, denoted as $\boldsymbol{v}^i_{l,border,adj}$ and $\boldsymbol{v}^i_{u,border,adj}$, at a given time instant $t_0 + \Delta t$, are referred to as PWCF and represent the output of the method, weher Δt is the sampling interval and t_0 is the actual time.

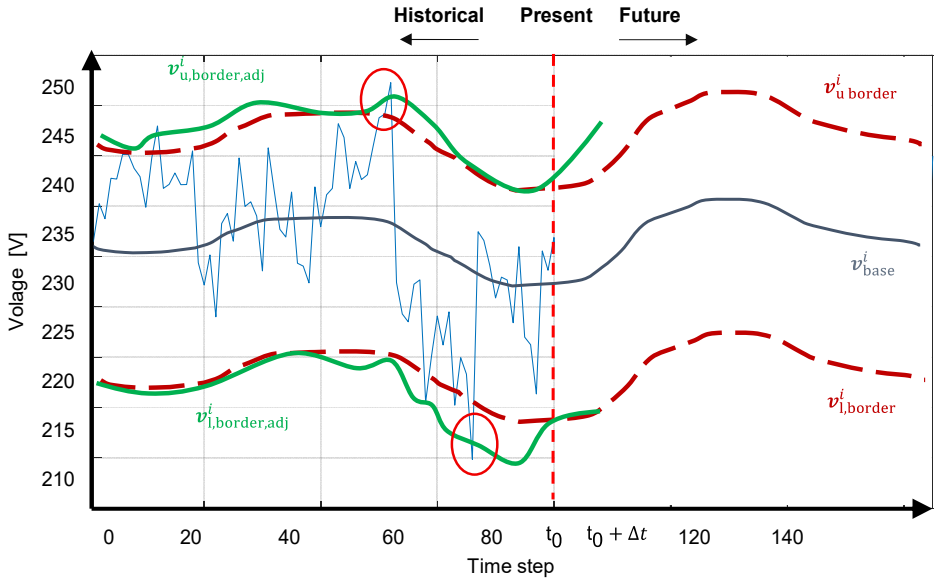


Figure 4.3: Visualization of the basic scheme of the PWCF method

However, it cannot be guaranteed that the PWCF will consistently remain within the defined secure operational limits. Ensuring complete avoidance of deviations would require significantly widening the forecast boundaries, which in turn would introduce an unnecessarily large safety margin and reduce the effectiveness of the control strategy.

4.2.1 Simple Pseudo-Worst-Case Forecasting Method

The voltages of each node and the branch currents at each sampling time are the recorded data. In the conventional forecasting method, the voltages and currents of the next sampling time are the outputs. But, in this study, the forecasted highest and lowest voltage values (and highest values of branch current) of the undesired conditions are considered as outputs. To prepare the input data, the matrix $\mathbf{V}_{his}^{i}$, for each controlled variable i is introduced in Eq. (4.1). This matrix is the same for the voltage of each node and branch current, and it contains historical information about each controlled variable i $(i \in \{1,2,..,N_c\}, N_c \in \mathbb{N}$, number of controlled variables N_c depends on the grid topology and the available actuators) at each sampling time in one day ($H = \frac{24 \cdot 60\ min}{R_s}$ which R_s Is the rate of sampling time of the meter). D is the number of days considered in the historical data.

$$\mathbf{V}_{his}^{i} = \begin{bmatrix} \mathrm{V}_{1,1}^{i} & \mathrm{V}_{1,2}^{i} & \cdots & \mathrm{V}_{1,H}^{i} \\ \mathrm{V}_{2,1}^{i} & \mathrm{V}_{2,2}^{i} & \cdots & \mathrm{V}_{2,H}^{i} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{V}_{D,1}^{i} & \mathrm{V}_{D,2}^{i} & \cdots & \mathrm{V}_{D,H}^{i} \end{bmatrix} \in \mathrm{N}^{D \times H} \tag{4.1}$$

In a simple case, these are only directly measured values from specific measurement devices, e.g., the voltage data from the smart meters. In a more advanced case, all available data are used to run a state estimation. The result is a state vector, i.e., all (phase-selective) node voltages and line currents are known (considering an estimation error). Nevertheless, not all values are relevant for the control, e.g., only the currents in the main lines are important, but not the currents in the house connection cables.

Several factors affect the values in each sample interval, including time (day, night, working hours), weather, season, holidays, and other factors. It is assumed that consumers more or less behave in similar patterns under the same conditions. The forecasting method is based on the assumption that the values of the controlled variables generally follow a daily pattern. This daily pattern, or profile, can be reproduced using historical data from a specific period. To create a dynamic forecast, the method also considers the dynamic behavior of the variable from the last four measurement cycles.

The first step is to reproduce the characteristic daily pattern from the historical data. To consider weather and season effects, the historical data for the D past days are selected. The matrix $\mathbf{V}_{his}^{i}$can be rearranged as the following Eq. (4.2):

$$\mathbf{V}_{his}^{i} = \begin{bmatrix} \boldsymbol{v}_{D,1}^{i} & \boldsymbol{v}_{D,2}^{i} & \cdots & \boldsymbol{v}_{D,H}^{i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_{H,1}^{i} \\ \boldsymbol{v}_{H,2}^{i} \\ \vdots \\ \boldsymbol{v}_{H,D}^{i} \end{bmatrix} \tag{4.2}$$

$\boldsymbol{v}_{H,D}^{i}$ refers to the sample vector, including the same sample for all D days. For example, all the sampling data at 8:00 for the D past days. $\boldsymbol{v}_{H,D}^{i}$ represents the most recent daily profile of the variable, containing the data from the last H sampling time (e.g., H in this study is 24 hours).

To define the base daily profile $\boldsymbol{v}_{base}^{i}$ the mean of all historical data at each time point is calculated using the following Eq.(4.3):

$$\boldsymbol{v}_{base}^{i}=\left[\text{mean}\left(\boldsymbol{v}_{D,1}^{i}\right) \quad \dots \quad \text{mean}\left(\boldsymbol{v}_{D,H}^{i}\right)\right] \in \mathbb{R}^{1\times H} \tag{4.3}$$

This base profile reflects the fundamental daily behavior of the variable. The observed data typically show high fluctuations with high stochastic gradients; therefore, for some purposes and depending on the number of pattern data used, the profiles are smoothed out with a smoother method in Eq. (4.4). K is the number of neighbors, and t_0 refers to the present time.

$$smooth: \quad \overline{\mathrm{V}}(t_0)=\frac{1}{2\cdot K+1}(\mathrm{V}(t_0-K)+\cdots+\mathrm{V}(t_0+K) \tag{4.4}$$

To smooth out any outliers, a moving average $\boldsymbol{v}_{base}^{i}$ is computed using a smoothing operator over five data points ($K=2$) as it is shown in Eq. (4.5):

$$\overline{\boldsymbol{v}}_{base}^{i}=smooth\left(\boldsymbol{v}_{base}^{i}\right) \in \mathbb{R}^{1\times H} \tag{4.5}$$

Next, for each time point, the maximum and minimum values are identified from the historical data. The matrix $\mathbf{V}_{his}^{i}$ is sorted so that the most significant values for each time point are placed in the first row, and the smallest values are placed in the last row as Eqs. (4.6), (4.7), (4.8), and (4.9).

$$\mathbf{V}_{\mathbf{s}}^{i}=\text{sort}\left(\mathbf{V}_{his}^{i}\right)=\begin{bmatrix} \widetilde{\mathrm{V}}_{1,1}^{i} & \widetilde{\mathrm{V}}_{1,2}^{i} & \dots & \widetilde{\mathrm{V}}_{1,\mathrm{H}}^{i} \\ \widetilde{\mathrm{V}}_{2,1}^{i} & \widetilde{\mathrm{V}}_{2,2}^{i} & \dots & \widetilde{\mathrm{V}}_{2,\mathrm{H}}^{i} \\ \vdots & \vdots & \cdots & \\ \vdots & \vdots & \cdots & \\ \widetilde{\mathrm{V}}_{\mathrm{D},1}^{i} & \widetilde{\mathrm{V}}_{\mathrm{D},2}^{i} & \dots & \widetilde{\mathrm{V}}_{\mathrm{D},\mathrm{H}}^{i} \end{bmatrix} \tag{4.6}$$

$$\left\{\widetilde{\mathrm{V}}_{d,h}^{i} \in \boldsymbol{v}_{D,t} \middle| \widetilde{\mathrm{V}}_{1,h}^{i} \geq \widetilde{\mathrm{V}}_{2,h}^{i} \geq \dots \geq \widetilde{\mathrm{V}}_{D,H}^{i}\right\} \in \mathbf{V}_{\mathrm{s}}^{i} \tag{4.7}$$

$$\boldsymbol{v}_{\max}^{i}=\left[\widetilde{\mathrm{V}}_{1,1}^{i} \quad \widetilde{\mathrm{V}}_{1,2}^{i} \quad \dots \quad \widetilde{\mathrm{V}}_{1,H}^{i}\right] \tag{4.8}$$

$$\boldsymbol{v}_{\min}^{\mathrm{i}}=\left[\widetilde{\mathrm{V}}_{D,1}^{i} \quad \widetilde{\mathrm{V}}_{D,2}^{i} \quad \dots \quad \widetilde{\mathrm{V}}_{D,H}^{i}\right] \tag{4.9}$$

Where $\boldsymbol{v}_{max}^{i}$ and $\boldsymbol{v}_{min}^{i}$ are respectively the vectors of maximum and minimum values for each time point across all days.

The smoothed base profile $\overline{\boldsymbol{v}}_{base}^{i}$ is used to create a forecast range by shifting the profile upward and downward by specific offsets. These offsets are determined by comparing the current day's basic pattern $\boldsymbol{v}_{base}^{i}$ with maximum and minimum values at the sampling time during the historical data, as shown in Eqs (4.10) and (4.11).

$$\boldsymbol{v}^{i}_{u,\,border}=\overline{\boldsymbol{v}}^{i}_{base}+((mean(\boldsymbol{v}^{i}_{max})-mean(\boldsymbol{v}^{i}_{base}))\cdot\mathbf{1}_{H\times1} \tag{4.10}$$

$$\boldsymbol{v}^{i}_{\mathrm{l},\,\mathrm{border}}=\overline{\boldsymbol{v}}^{i}_{base}+((mean(\boldsymbol{v}^{i}_{min})-mean(\boldsymbol{v}^{i}_{base}))\cdot\mathbf{1}_{H\times1} \tag{4.11}$$

$\mathbf{1}_{H\times1}$ is the unity vector where every element is 1. Note that in each sample only elements at time ($t_0+\Delta t$) in vectors $\boldsymbol{v}^{i}_{u,border}$and $\boldsymbol{v}^{i}_{l,border}$ are important, which represent the values for the next sample.

To account for the dynamic behavior of the grid, a mechanism referred to as “dynamic adjustment” is introduced. This approach comprises two components: the long-term adjusted value ($\mathrm{V}^{i}_{lt,adj}$) Eq.(4.12), and the short-term adjusted value($\mathrm{V}^{i}_{st,adj}$), Eq. (4.13), each addressing different temporal aspects of system variability. Long-term adjustments take into account the past six sampling times. For example, if the sampling rate is 10 minutes, this long-term consideration reflects consumer behavior over the past hour. The idea is that changes occurring in the last hour, such as shifts in weather or consumer behavior, can still have an impact on the present time. Short-term adjustments refer to sudden changes that occurred recently and may still influence the current values. In both cases, the adjusted value is the difference between the mean value of the last samples and the smoothed base pattern, built for different numbers of samples.

$$\begin{aligned}
&V^{i}_{lt,adj}=mean(\boldsymbol{v}^{i}_{H,lt})-mean(\overline{\boldsymbol{v}}^{i}_{base,lt}) \quad \text{with}\\
&\boldsymbol{v}^{i}_{H,lt}=\left\{\mathrm{V}^{i}_{1,h}\in\mathbf{V}^{i}_{his}\middle|h\in\{1,\dots,6\}\right\}\\
&\overline{\boldsymbol{v}}^{i}_{base,lt}=\left\{\overline{\mathrm{V}}^{i}_{base,h}\in\overline{\boldsymbol{v}}^{i}_{base}\middle|h\in\{1,\dots,6\}\right\}
\end{aligned} \tag{4.12}$$

$$\begin{aligned}
&\mathrm{V}^{i}_{st,adj}=mean(\boldsymbol{v}^{i}_{H,st})-mean(\overline{\boldsymbol{v}}^{i}_{base,st}) \ \text{ with}\\
&\boldsymbol{v}^{i}_{H,st}=\left\{\mathrm{V}^{i}_{1,h}\in\mathbf{V}^{i}_{his}\middle|h\in\{1,2\}\right\}\\
&\overline{\boldsymbol{v}}^{i}_{base,st}=\left\{\overline{\mathrm{V}}^{i}_{base,h}\in\overline{\boldsymbol{v}}^{i}_{base}\middle|h\in\{1,2\}\right\}
\end{aligned} \tag{4.13}$$

The adjusted values are used to make new terms to adjust the pseudo-worst-case lower and upper borders according to Eq. (4.14) and (4.15). As shown in Figure 4.4, the red dashed lines are now adjusted to green lines $\mathrm{V}^{i}_{u,border,adj}$, and $\mathrm{V}^{i}_{l,border,adj}$. The adjusted borders for a certain value i are set by selecting the minimum (in case of $\mathrm{V}^{i}_{l,border,adj}$.) or the maximum (in case of $\mathrm{V}^{i}_{u,border,adj}$) of four different terms.

$$\boldsymbol{v}_{u,fc}=\begin{bmatrix}\dots & \mathrm{V}^{i}_{u,border,adj} & \dots\end{bmatrix}^{T} \tag{4.14}$$

$$\begin{aligned}
\mathrm{V}^{i}_{u,border,adj}=\max(&\mathrm{V}^{i}_{u,border}\,,\mathrm{V}^{i}_{u,border}+mean(\mathrm{V}^{i}_{lt,adj},\mathrm{V}^{i}_{st,adj})\,,\mathrm{V}^{i}_{u,border}\\
&+\mathrm{V}^{i}_{lt,adj}\,,\mathrm{V}^{i}_{1,1},\mathrm{V}^{i}_{1,2})
\end{aligned} \tag{4.15}$$

$$\boldsymbol{v}_{l,fc} = \left[\ldots \quad \mathrm{V}^{i}_{l,border,adj} \quad \ldots\right]^{T} \tag{4.16}$$

$$\mathrm{V}^{i}_{l,border,adj} = \min(\mathrm{V}^{i}_{l,border}, \mathrm{V}^{i}_{l,border} + mean\left(\mathrm{V}^{i}_{lt,adj}, \mathrm{V}^{i}_{st,adj}\right), \mathrm{V}^{i}_{l,border} + \mathrm{V}^{i}_{lt,adj}, \mathrm{V}^{i}_{1,1}, \mathrm{V}^{i}_{1,2}) \tag{4.17}$$

For example, $\mathrm{V}^{i}_{l,border,adj}$ in Eq.(4.16). is the minimum value out of the following terms: First, the lower border $\mathrm{V}^{i}_{l,border}$. Second, the lower border shifted down with the mean value of the long-term and short-term adjusted values to consider high changes in the long-term behaviour, moderated by the recent behaviour (short term). Third, the lower border shifted down with the amount of the long-term adjusted value. Lastly, the present or previous value leads to the fact that the lower pseudo-worst-case value forecast is always smaller than or equal to these two values. Eq. (4.17) constitutes a heuristic formulation derived through systematic trial-and-error testing with the available dataset (see subchapter 5.1.1 for details). This expression was established by evaluating alternative terms and selecting those that consistently yield the most reliable results for different dynamic behaviors of the observed values.

In [141], the concept of the PWCF is employed to establish a synthetic range, defined by upper and lower boundaries, within which a regulated parameter (e.g., voltage or current) is expected to reside during the upcoming control cycle. Rather than attempting to predict exact snapshot values, which are often highly volatile, the PWCF method integrates both historical and current data to construct a robust envelope. This approach enables grid automation systems to enact preventive measures when the forecasted worst-case scenario indicates that the parameter may exceed acceptable tolerance limits.

4.2.2 Forecasting with Neural Network Method

This chapter explores several methodologies utilizing NN for forecasting, focusing particularly on feedforward architectures and placing a strong emphasis on the dependency of these methods on the nature of inputs and outputs. A NN is a computational model inspired by the human brain, which was first introduced by McCulloch and Pitts (1943) [160]. The core of forecasting in NNs lies in the careful selection of inputs and outputs, which fundamentally dictate the effectiveness of the model. In the following methods, NNs are implemented with varying configurations of input and output variables. While the underlying NN modeling framework is well established in the literature, the contribution of this study resides in the systematic design and evaluation of alternative input and output combinations. This process highlights the critical impact of data on the forecasting capabilities of NN. Initially, convolutional NN methods are briefly reviewed, their limitations highlighted, and subsequently, the improved PWCF method with the NN method is introduced.

4.2.2.1 Forecasting Utilizing Present and Next Step Data

In the simplest method examined within NN-based forecasting approaches, the present data

serves as the input, while the subsequent time step data is considered the output. This method, which is shown in Figure 4.4, is fundamental in its approach, selects data observed at a current time t_0 as the input to the model. The outputs are the predictions for the next time step, $t_0 + \Delta t$ represents the next sampling time. Eq. (4.18) outlines these inputs and outputs. The NN aims to find the best fitting pattern between these input and output values by using a subset of historical data for training.

$$\boldsymbol{v}_{in,t_0} = \begin{bmatrix} \mathrm{V}_{t_0}^1 \\ \mathrm{V}_{t_0}^2 \\ \\ \mathrm{V}_{t_0}^N \end{bmatrix} \qquad \boldsymbol{v}_{out,t_0+\Delta t} = \begin{bmatrix} \mathrm{V}_{t_0+\Delta t}^1 \\ \mathrm{V}_{t_0+\Delta t}^2 \\ \\ \mathrm{V}_{t_0+\Delta t}^N \end{bmatrix} \tag{4.18}$$

Where V_{i,t_0}represents the data of node i at time t_0, which is the present time and $t_0 + \Delta t$ is the next sampling time. $\mathrm{V}_{1,t_0}, \mathrm{V}_{1,t_0+\Delta t}$ are the data at present and the predicted data of the sampling time, respectively, and $\mathrm{n}=1, 2,\ldots, \mathrm{N}$ denotes the number of nodes.

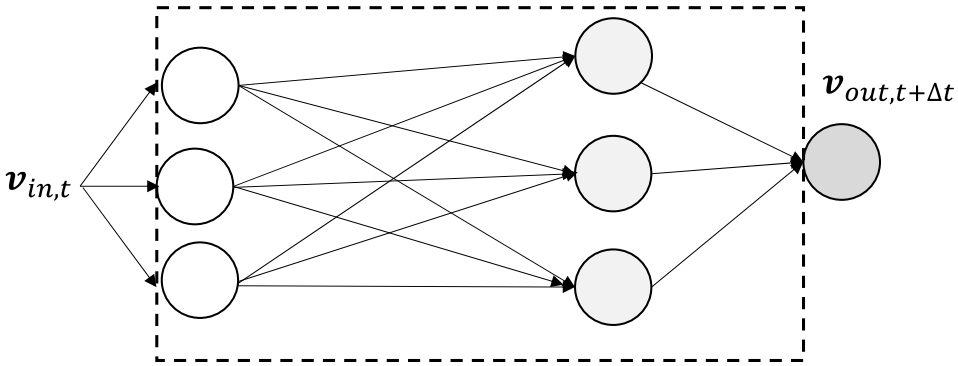


Figure 4.4: Present and next step data for the NN method

4.2.2.2 Clustering Input and Output Data

To enhance forecasting accuracy, this method uses data clustering by classifying the variables according to their specific operational limits, such as the maximum and minimum permissible voltage levels defined for low-voltage grids. In accordance with EN 50160, the voltage at the customer connection point is generally allowed to vary within ±10% of the nominal voltage. In the following sections, the equations and explanations are provided using voltage as an example. Data are classified into six groups based on nominal values, as shown in Eq.(4.19).

$$\begin{cases} \mathrm{V}_{nom} \leq \mathrm{V}_{i,n} \leq (1.05)\mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_1 \\ (1.05)\mathrm{V}_{nom} \leq \mathrm{V}_{i,n} \leq (1.1)\mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_2 \\ \mathrm{V}_{i,n} \geq (1.1)\mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_3 \\ (0.95)\mathrm{V}_{nom} \leq \mathrm{V}_{i,n} \leq \mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_4 \\ (0.9)\mathrm{V}_{nom} \leq \mathrm{V}_{i,n} \leq (0.95)\mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_5 \\ \mathrm{V}_{i,n} \leq (0.9)\mathrm{V}_{nom} & \mathrm{V}_{i,n} \in \mathrm{C}_6 \end{cases} \tag{4.19}$$

The clustering approach operates according to permissible voltage boundaries (±10% of the nominal voltage). Two clusters are near V_{nom} (±5% around V_{nom}), two are near the permissible

lower and upper boundaries, and two represent the worst-case scenarios outside the permissible range. For each node, the input data are the clustered groups, and the outputs are the data at the next sampling time as defined in Eq. (4.20) and shown in Figure 4.5.

$$\boldsymbol{v}_{in\_clustered,t_0} = \begin{bmatrix} \mathrm{V}_{C,t_0}^{1} \\ \mathrm{V}_{C,t_0}^{2} \\ \vdots \\ \mathrm{V}_{C,t_0}^{N} \end{bmatrix} \qquad \boldsymbol{v}_{out,t_0+\Delta t} = \begin{bmatrix} \mathrm{V}_{t_0+\Delta t}^{1} \\ \mathrm{V}_{t_0+\Delta t}^{2} \\ \vdots \\ \mathrm{V}_{t_0+\Delta t}^{N} \end{bmatrix} \tag{4.20}$$

Where V_{C,t_0}^{i} represents the clustered data of node i at time t_0.

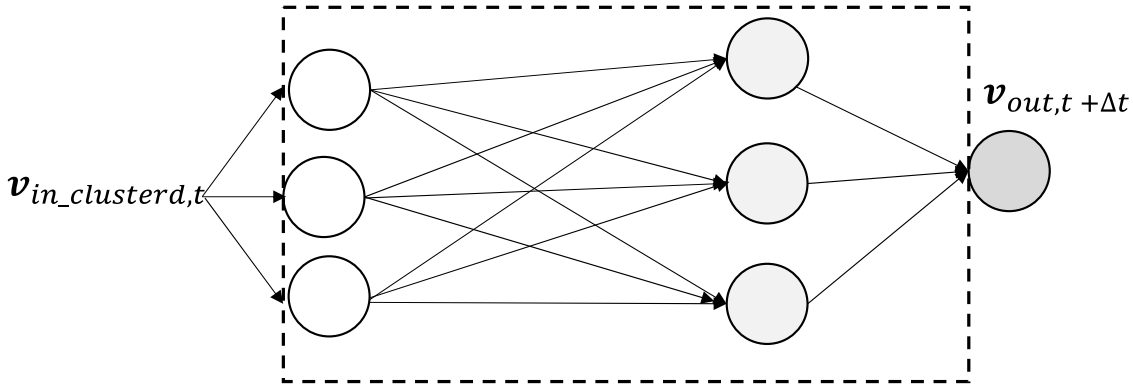


Figure 4.5: Clustering input and output data

4.2.2.3 Inputting Differences from Real Data

This method focuses on the next step of data forecasting based on deviations from a base pattern of behavior. This method is influenced by various factors such as time, weather, and whether it is a weekday or a weekend. The base pattern, which is introduced in Eq. (4.3), $\boldsymbol{v}_{base}^{i}$, represents the expected behavior at each node. The real data at time t_0 ($\mathrm{V}_{t_0}^{i}$) and deviations between the actual observed data ($\mathrm{V}_{t_0}^{i}$) and this base pattern at time t_0 ($\boldsymbol{v}_{base}^{i}$), as it is explained in Eq. (4.21), are used as input data and the deviated data in the next sampling time, which is introduced in Eq. (4.22), is considered as output. The inputs and output vectors are presented in Eq. (4.23-4.25) and shown in Figure 4.6.

$$\mathrm{V}_{diff,t_0}^{i} = \mathrm{V}_{t_0}^{i} - \mathrm{V}_{base,t_0}^{i} \tag{4.21}$$

$$\mathrm{V}_{out,t_0+\Delta t}^{i} = \mathrm{V}_{diff,t_0}^{i} + \mathrm{V}_{base,t_0+\Delta t}^{i} \tag{4.22}$$

$$\boldsymbol{v}_{in,t_0} = \begin{bmatrix} \mathrm{V}_{t_0}^{1} \\ \mathrm{V}_{t_0}^{2} \\ \\ \mathrm{V}_{t_0}^{N} \end{bmatrix} \; and \; \boldsymbol{v}_{diff,t_0} = \begin{bmatrix} \mathrm{V}_{diff,t_0}^{1} \\ \mathrm{V}_{diff,t_0}^{2} \\ \vdots \\ \mathrm{V}_{diff,t_0}^{N} \end{bmatrix} \qquad \boldsymbol{v}_{out,t_0+\Delta t} = \begin{bmatrix} \mathrm{V}_{out,t_0+\Delta t}^{1} \\ \mathrm{V}_{out,t_0+\Delta t}^{2} \\ \vdots \\ \mathrm{V}_{out,t_0+\Delta t}^{N} \end{bmatrix} \tag{4.23-4.25}$$

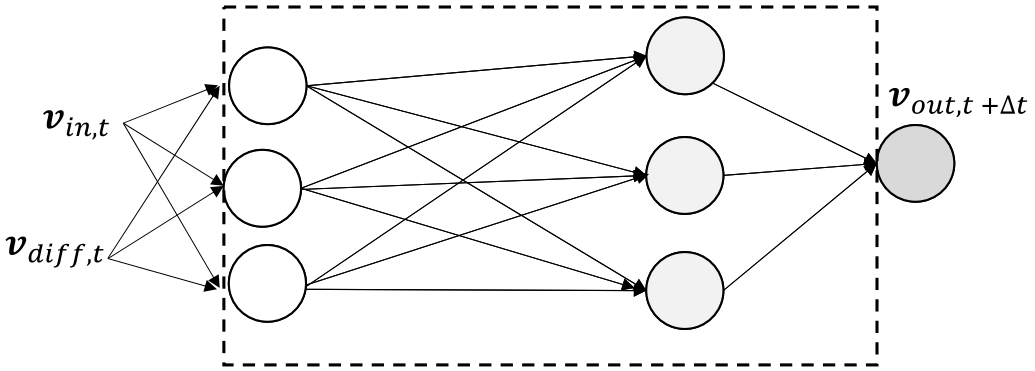


Figure 4.6: Inputting differences from real data

One major challenge with these methods (4.2.2.1, 4.2.2.2, and 4.2.2.3) is dealing with the highly fluctuating real-world data. The model often struggles to learn from sudden changes in the dataset, which are critical for making accurate forecasts. To capture these rapid fluctuations, the model requires not only historical data but also a sophisticated mechanism to identify and learn from these anomalies. For this reason, instead of relying solely on precise point forecasting using NN methods, where the model attempts to predict the next-step value, a hybrid approach is introduced. In this approach, the NN models are combined with the PWCF framework to capture uncertainty and variability better. This combined method is presented in detail in the following chapter.

4.2.3 Improving the Pseudo-Worst Case Forecasting Method by Using the Neural Network Method

Given the challenges identified with previous methods, particularly their struggle to adapt to the unpredictable nature of real-world data, the focus has shifted to a PWCF method. In practical applications, forecasting is required to support system operators by providing early indications of potential worst-case situations and sudden changes. Therefore, the novel heuristic PWCF approach is enhanced by integrating an established NN model, with the aim of more reliably anticipating critical events through improved mapping between selected input and output variables.

The method presented in this section further develops the previously introduced PWCF method. In the earlier version, the technique forecasted upper and lower worst-case boundaries for voltage and branch current at each node in the grid, representing conditions that might occur at the next time step. These boundaries were determined as the maximum and minimum values among probable dynamic behavior scenarios, which were chosen based on prior expert knowledge of the grid behavior. In contrast, the current method uses an NN to forecast the upper and lower boundaries, making the forecasting process independent of operator intervention. This approach leverages historical data, similar to the previous method, but utilizes an NN to refine forecasting. Here, the inputs are the present data of individual nodes, and the outputs are the predictions for the next sampling time.

Similar to the PWCF method introduced in section 4.2.1, this approach utilizes a historical data matrix $\mathbf{V}_{his}^{i}$, which is introduced in Eq. (4.2), to forecast the voltage and current of each node. This matrix captures the daily basic pattern ($\boldsymbol{v}_{base}^{i}$), which is chosen as a 28-day period to ensure consistency in weather and seasonal influences. Deviations from this base pattern $\mathrm{V}_{diff,t_0}^{i}$ are then calculated, as indicated in Eq. (4.21), representing the differences between the actual data at time t_0 and the base pattern ($\boldsymbol{v}_{base}^{i}$) (Eq. (4.8)). These deviations are crucial for identifying shifts due to external influences such as weather changes. To address the potential for extreme cases, the historical data matrix is sorted from maximum to minimum values to identify the worst-case scenarios historically observed; the sorted matrix ($\mathbf{V}_{s}^{i}$) is introduced in Eqs. (4.6). From this sorted matrix, the maximum and minimum values are calculated, $\mathrm{V}_{max,t_0}^{i} \in \boldsymbol{v}_{max}^{i}$ as Eq. (4.7), and $\mathrm{V}_{min,t_0}^{i} \in \boldsymbol{v}_{min}^{i}$ Eq. (4.8). The upper difference value ($\mathrm{V}_{diff,max,t_0}^{i}$) and the lower difference value ($\mathrm{V}_{diff,min,t_0}^{i}$) are presented in Eq. (4.24) and (4.25), which represent the potential upper and lower bounds for future data points. These calculations guide the determination of critical boundaries for forecasting, particularly under worst-case conditions.

$$\mathrm{V}_{diff,max,t_0}^{i} = \mathrm{V}_{i,t_0} - \mathrm{V}_{max,t_0}^{i} \tag{4.26}$$

$$\mathrm{V}_{diff,min,t_0}^{i} = \mathrm{V}_{i,t_0} - \mathrm{V}_{min,t_0}^{i} \tag{4.27}$$

To find the upper and lower critical borders ($\mathrm{V}_{u,\,border}^{i}$ and $\mathrm{V}_{l,border}^{i}$) that are shown with red dashed lines in Figure 4.3, $\mathrm{V}_{base,t_0+\Delta t}^{i}$, the basic value at time $t_0 + \Delta t$, is shifted by an offset (absolute value of the difference value $\mathrm{V}_{diff,t_0}^{i}$) to the up and down to create the upper and lower values at time $t_0 + \Delta t$.

$$\mathrm{V}_{u,border,t_0+\Delta t}^{i} = \mathrm{V}_{base,t_0+\Delta t}^{i} + \left|\mathrm{V}_{diff,t_0}^{i}\right| \tag{4.28}$$

$$\mathrm{V}_{l,border,t_0+\Delta}^{it} = \mathrm{V}_{base,t_0+\Delta t}^{i} - \left|\mathrm{V}_{diff,t_0}^{i}\right| \tag{4.29}$$

In addressing the dynamic behavior of household data, this method considers the inherent uncertainties by introducing adjustments to the forecasted boundaries based on potential extreme variations. These adjustments, defined as $\mathrm{d}_{worst,max,t_0+\Delta t}^{i}$ and $\mathrm{d}_{worst,min,t_0+\Delta t}^{i}$ or the respective maximum and minimum deviations, are critical for ensuring the resilience of the model against unexpected fluctuations. These deviations are added to the previously calculated borders ($\mathrm{V}_{u,border,t_0+\Delta t}^{i}$ and $\mathrm{V}_{l,border,t_0+\Delta t}^{i}$) to encompass a broader range of potential outcomes, ensuring robustness in the forecasting method. The adjusted upper and lower borders are detailed in Eq. (4.30) and Eq. (4.31).

$$\mathrm{V}_{u,border,adj,t_0+\Delta t}^{i} = \mathrm{V}_{u,border,t_0+\Delta t}^{i} + \mathrm{d}_{worst,max,t_0+\Delta t}^{i} \tag{4.30}$$

$$\mathrm{V}_{l,border,adj,t_0+\Delta t}^{i} = \mathrm{V}_{l,border,t_0+\Delta t}^{i} - \mathrm{d}_{worst,min,t_0+\Delta t}^{i} \tag{4.31}$$

These deviations are derived from the worst-case scenarios within the historical data, specifically from the maximum and minimum values observed, which are recalculated to consider recent changes that may affect future states. These recalculations, $d^{i}_{worst,max,t_0+\Delta t}$ and $d^{i}_{worst,min,t_0+\Delta t}$ are proportionally adjusted by the factors α_{max} and α_{min} to scale the impact of these extreme values on the forecasted ranges, as shown in Eqs. (4.32) and (4.33).

$$d^{i}_{worst,max,t_0+\Delta t}=\alpha_{max}\cdot V^{i}_{max,t_0+\Delta t} \tag{4.32}$$

$$d^{i}_{worst,min,t_0+\Delta t}=\alpha_{min}\cdot V^{i}_{min,t_0+\Delta t} \tag{4.33}$$

The values of α_{max} and α_{min} determine how much the upper and lower forecasting boundaries should be expanded. These values are not chosen arbitrarily. Instead, they are calculated from the probability density function (PDF) of the historical deviations. The PDF describes how likely each deviation value is. Large deviations that rarely occur in the past correspond to small PDF values, while small, frequent deviations correspond to large PDF values.

By using the PDF, the model can distinguish between “normal” behavior and “extreme” behavior. Extreme deviations receive larger adjustment factors because they indicate a higher risk that the next data point may exceed the normal range. In contrast, common deviations receive smaller adjustments. Eq. (4.34) uses this idea and converts the deviation into the two adjustment factors α_{max} and α_{min}. It does this through a piecewise rule:

- if the present deviation is positive, only the upper boundary receives a large adjustment,
- if the present deviation is negative, only the lower boundary receives a large adjustment.

In this way, Eq. (4.34) directly connects the statistical information from the PDF with the dynamic adjustment of the prediction boundaries.

$$\begin{aligned}
&\alpha_{max}=f(V^{i}_{diff,max,t_0}) && \text{if } V^{i}_{diff,t_0}>0\\
&\alpha_{min}=f_{max}-f(V^{i}_{diff,max,t_0}) && \text{if } V^{i}_{diff,t_0}>0\\
&\alpha_{max}=f(V^{i}_{diff,min,t_0}) && \text{if } V^{i}_{diff,t_0}\leq 0\\
&\alpha_{min}=f_{max}-f(V^{i}_{diff,min,t_0}) && \text{if } V^{i}_{diff,t_0}\leq 0
\end{aligned} \tag{4.34}$$

This probabilistic approach, shown in Eq. (4.35), quantifies the uncertainty and adjusts the forecasting boundaries accordingly. The quantity $f_{max}=\max_{x\in H} f(x)$ is defined as the maximum PDF value within the historical deviation set H.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4.35}$$

The mean of the deviation distribution is set to $\mu=0$ because each deviation V^{i}_{diff} is computed as the difference between the real data and its corresponding base pattern. Since the base pattern represents the long-term average behavior of each node, the resulting deviations oscillate around zero. The standard deviation is set to $\sigma=30$ to create a sufficiently wide PDF,

ensuring that both normal and extreme deviations receive distinct probability weights. This statistical framework is visually represented in Figure 4,7, which illustrates how the factors α_{max} and α_{min}. are selected based on the PDF, linking statistical analysis directly to practical forecasting applications. This approach not only enhances the accuracy of forecasting but also ensures that they are robust enough to handle the unpredictable nature of household energy consumption patterns effectively.

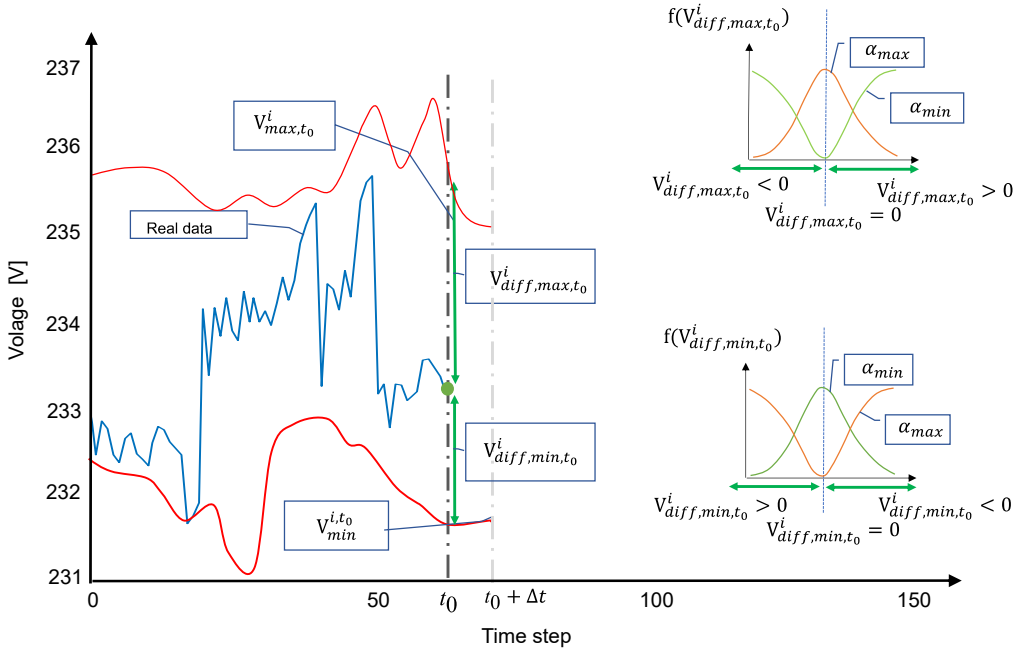


Figure 4.7: Factor selection with PDF diagrams

Clustering is performed in three distinct zones. In the following, each zone is explained: Zone 1 includes cases where the real data lie within the region defined by the permissible borders. Zone 2 encompasses cases where the real data fall outside this region. Finally, Zone 3 corresponds to cases where the real data lie exactly on the permissible borders.

Zone 1: $V^i_{diff,max,t_0} < 0$ or $V^i_{diff,min,t_0} > 0$

The data resides within the permissible borders region, closely aligned with the permissible borders. In such instances, the proximity of data to these borders reduces the differential, which consequently heightens the likelihood of encountering a worst-case scenario. For instance, when $V^i_{diff,max,t_0} \cong 0$, there exists a substantial probability that subsequent measurements could exceed these boundaries. Figure 4.7 illustrates this condition, showing that under these circumstances, α_{max}reaches its peak, implying a minimal chance that data will fall below the lower limit.

Zone 2: $V^{i,}_{diff,max t_0} > 0$ or $V^i_{diff,min,t_0} < 0$

Here, the data points either exceed the upper permissible boundary ($V^i_{diff,max,t_0} > 0$) or

fall below the lower boundary ($V^{i}_{diff,min,t_0} < 0$). This scenario is akin to Zone 1, where proximity to the threshold has a direct influence on the likelihood of breaching these limits. Notably, if data trends towards these extremities, such as when V^{i}_{diff,max,t_0} or V^{i}_{diff,min,t_0} is near zero, the corresponding α values adjust to their extremes to reflect the increased risk of crossing these boundaries.

Zone 3: $V^{i}_{diff,max,t_0} = 0$ or $V^{i}_{diff,min,t_0} = 0$

This zone is characterized by data points exactly mirroring the historical extremes, either maximum or minimum. This exact match indicates a strong likelihood that upcoming data points will venture outside the established permissible margins, particularly in the next measurement interval. Such situations demand heightened vigilance and adaptive forecasting strategies to pre-emptively address potential anomalies.

Figure 4.8 illustrates the methodology applied to analyze voltage stability at each grid node during each sampling interval. The inputs for this analysis include the voltage at each node (V^{i,t_0}_{in}) at the initial time t_0, along with the maximum ($V^{i,t_0+\Delta t}_{max}$) and minimum ($V^{i,t_0+\Delta t}_{min}$) voltages recorded historically at the next time interval $t_0 + \Delta t$. Additionally, the difference between the current and expected voltage, (V^{i,t_0}_{diff}), is used to measure deviations. The outputs of the analysis are the adjusted worst-case scenario boundaries for the next sampling time. These boundaries, $V^{i,t_0+\Delta t}_{u,border,adj}$ for the upper limit and $V^{i,t_0+\Delta t}_{l,border,adj}$ for the lower limit, are crucial for pre-emptive voltage regulation. They help ensure network stability by accommodating potential voltage fluctuations beyond normal operational thresholds, thus enhancing the grid's reliability and safety.

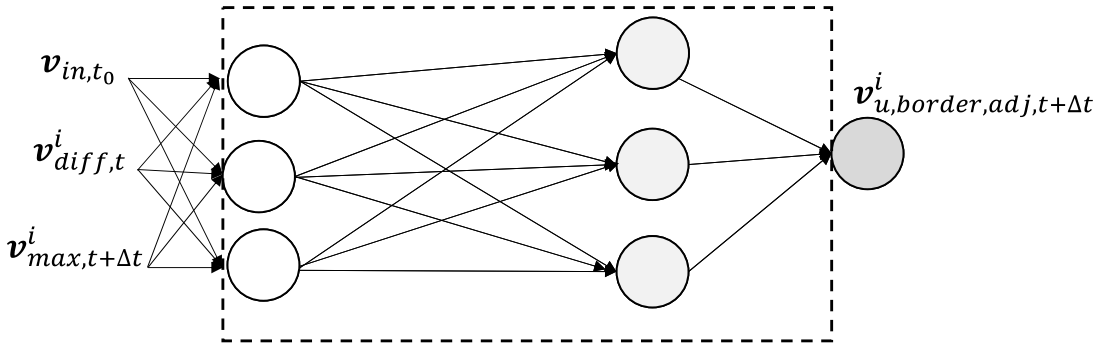


Figure 4.8: New approach of PWCF with NN

4.3 Anomaly Detection for Pattern Change Using Artificial Intelligence Methods

In data-driven and AI-based methodologies, the effectiveness of a model depends not only on the selection of appropriate algorithms and parameter tuning but also critically on the choice and preparation of input data. This thesis explores various methods and input-handling strategies to demonstrate how different combinations of models and input configurations can

be tailored to the available data and intended objectives. In some cases, raw input data may be used directly; in others, incorporating additional features or external data sources can significantly enhance performance. Decomposing input data to isolate specific patterns, such as separating trend and residual components, has also shown benefits. Moreover, structuring input as sequences, as done in Long Short-Term Memory (LSTM) Autoencoders, represents another effective strategy. These diverse techniques underscore the importance of thoughtful input design in improving both the robustness and accuracy of AI-driven models [142, 143].

The Isolation Forest method, first introduced by Liu, Ting, and Zhou in 2008 [144], is an unsupervised anomaly-detection algorithm based on the idea that anomaly samples can be isolated more quickly than normal data points. Throughout this section, several of the proposed approaches build upon this foundational concept.

In summary, although the ML algorithms used in this thesis are established methods from the literature and implemented using standard Python libraries, the way they are selected, prepared, and combined with statistical techniques is developed in this work. In particular, the design of input configurations, the use of statistical features, and the integration of different models are proposed as part of this research.

4.3.1 Statistical Windowing with Isolation Forest

The isolation Forest method is a highly efficient and scalable approach for anomaly detection, valued for its simplicity and minimal need for parameter tuning, primarily requiring the specification of the number of trees and sub-sample size. This makes it particularly well suited for large datasets, as it enables rapid detection of anomalies without imposing a significant computational load. The core Isolation Forest algorithm is implemented in this study following the original framework described in [144]

As illustrated in Figure 4.9, this method involves carefully selecting specific input data for model training and evaluation. The inputs include extra data, such as global solar radiation and temperature, which provide additional contextual information relevant to the system's behavior, and extra features, such as the mean (μ_w) and standard deviation (σ_w) computed over defined time windows. These features are derived from the input data and help to capture short-term fluctuations and local patterns.

This input data is then passed to the isolation Forest model, which isolates potential outliers based on how easily they can be separated from the rest of the data points. The model's output is then post-processed to determine and label anomalous events. By combining statistical features with additional environmental or operational variables, this approach enables the detection of anomalies that may remain hidden when relying solely on raw data.

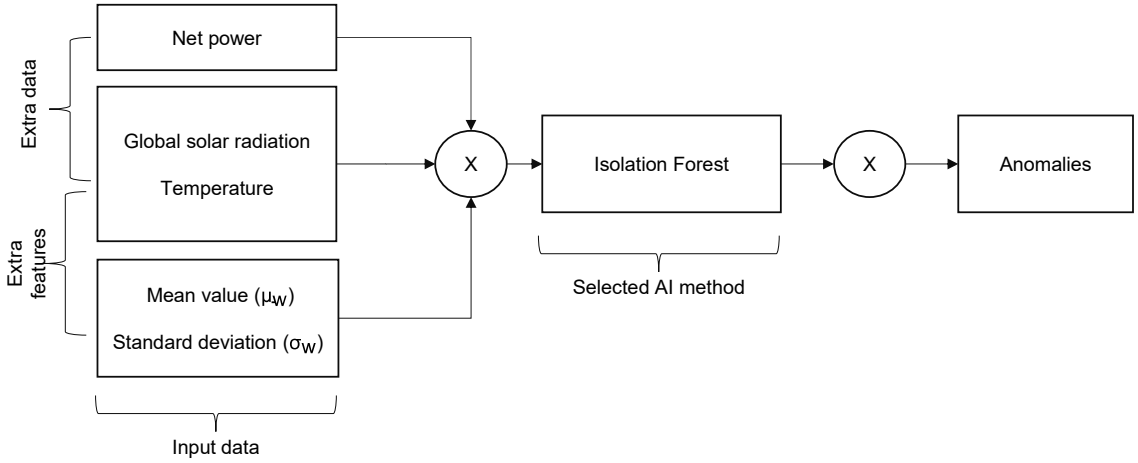


Figure 4.9: Structure of statistical windowing with the isolation Forest method

4.3.2 Cluster-Enhanced Feature Regression

Linear Regression, first introduced by Galton in 1894 [145, 146], is now widely used as one of the simplest machine-learning methods for predictive analysis of continuous variables and serving as the basis for numerous practical implementations in modern programming environments such as Python [147].

This algorithm divides the data into two categories based on solar potential: high power variability ($\mathbf{c}_{\mathrm{HPV}}$) and low power variability ($\mathbf{c}_{\mathrm{LPV}}$), determined by monthly sunshine probability. For months with a high probability of sunshine, the regression model uses global solar radiation and day of the week as input features, since these variables are strongly correlated with power output in such conditions. Conversely, during months with lower sunshine probability, the model relies on temperature and day of the week, which show a stronger correlation with power behavior in these cases. This dynamic feature selection enhances the model's ability to reflect seasonal and contextual changes. The day of the week remains a consistent input across both groups, as it reliably correlates with energy usage patterns in urban environments.

A linear regression model is then trained on the selected features to predict energy usage. Anomalies are detected by analyzing the prediction residuals, specifically, deviations from the predicted values that exceed a defined threshold based on the residuals' standard deviation. These significant deviations are flagged as anomalies. Figure 4.10 illustrates the overall structure of this approach, showing the selection of input features, the application of the regression model, and the subsequent detection of anomalies

4.3.3 Correlative Isolation Forest

The proposed algorithm uses the isolation Forest method to detect anomalies by focusing on comparing data under similar conditions. This is particularly useful when the goal is to detect changes while maintaining the same environmental context, ensuring that anomalies are identified relative to comparable conditions.

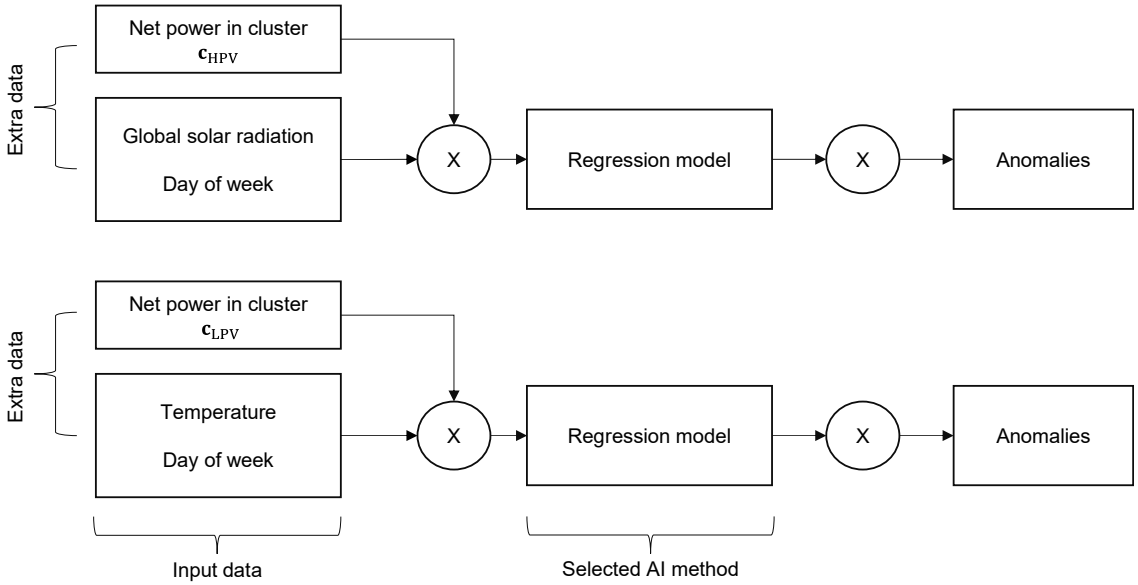


Figure 4.10: Structure of cluster-enhanced feature regression method

In the proposed algorithm, the complete dataset D is partitioned into $\mathrm{n_w}$ sequential windows, each referred to as $\mathrm{W_i}$, with length w. The methodology integrates key environmental variables such as temperature, wind, seasonal changes, and global solar radiation into a feature matrix $\mathbf{F}$, for correlation analysis. For each window $\mathrm{W_i}$, up to the $\mathrm{n_w}$-1 window, a correlation score is computed by averaging the correlation coefficients between the features in the target data window ($\mathbf{F}_{\mathrm{T}}$) and those in each respective window ($\mathbf{F}_{\mathrm{W_i}}$). The algorithm ranks all the windows based on their correlation scores and identifies the top k windows that exhibit the highest correlations. The data considered from this specific window are denoted as $\mathbf{y}_{\mathrm{Top_k}}$, are utilized to train an isolation Forest model.

This model detects anomalies by learning the patterns from the most correlated windows of data ($\mathbf{y}_{\mathrm{Top_k}}$). The model is then applied to the data of the selected window $\mathrm{W_T}$ ($\mathbf{y}_{\mathrm{W_T}}$) to ascertain the presence of anomalies. This correlation-based approach ensures that anomaly detection is both robust and context-sensitive, reducing the likelihood of false positives and increasing the relevance of the detected anomalies. The structure of inputs and outputs is shown in Figure 4.11.

4.3.4 Hybrid Extreme Studentized Deviate

The proposed algorithm combines several powerful methods to detect even the smallest anomalies in time series data, enhancing its sensitivity. It combines Seasonal and Trend decomposition using Loess (STL) (the method will be explained in detail in 4.3.5) with the Extreme Studentized Deviate (ESD) test, a well-established statistical method for identifying outliers [148]. STL effectively decomposes time series data into three distinct components: the trend (the long-term progression or direction of data), the seasonal component (recurring fluctuations related to calendar or periodic cycles), and the residual (the remaining variations

after removing trend and seasonality) ($\mathbf{y}_i = \mathbf{y}_{trend_i} + \mathbf{y}_{s_i} + \mathbf{y}_{res_i}$).

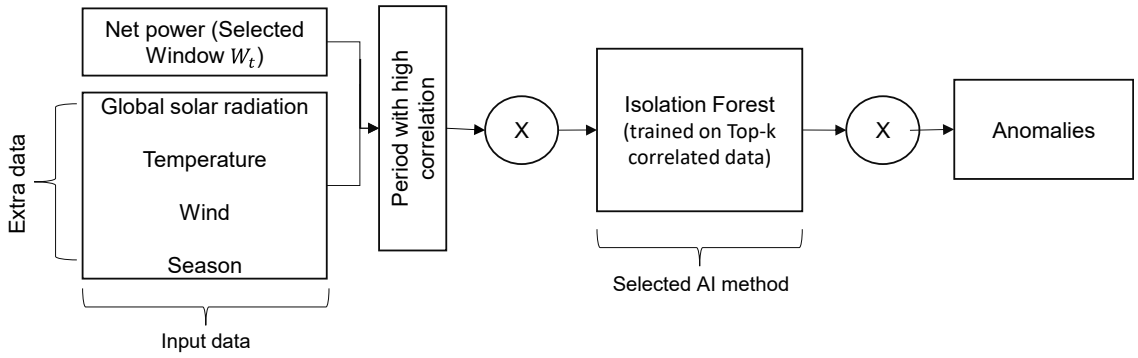


Figure 4.11: Structure of the correlation isolation Forest method

This decomposition clarifies underlying patterns by isolating and analyzing each component separately. Subsequently, the isolation Forest algorithm is trained using the residual component. Anomalies are at the point where the standardized scale ($\mathbf{Z}_{res_i} = \frac{\mathbf{y}_{res_i} - \bar{\mathbf{y}}_{res}}{\sigma_{res}}$) of the residual data is larger than the ESD factor, which is introduced in [148] ($\lambda_k = \frac{(n-k)\cdot t_{\alpha/(2(n-k)),n-k}}{\sqrt{(n-k)^2 + t^2_{\alpha/(2(n-k)),n-k}}}$).

Here, $\bar{\mathbf{y}}_{res}$ is the mean of the residuals vector data, σ_{res} is the standard deviation of the residuals, n is the total number of observations, k is the current observation, t is the critical value from the t-distribution for the given confidence level α and degrees of freedom n-k. The method is shown in Figure 4.12.

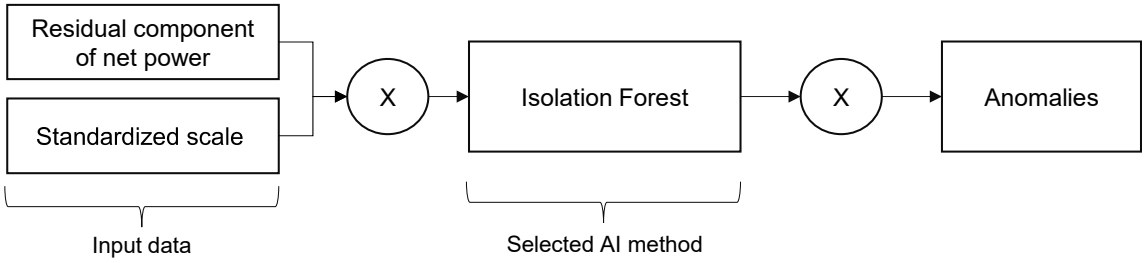


Figure 4.12: Structure of the hybrid ESD method

4.3.5 Seasonal and Trend Decomposition Using Loess Method

This method is a combination of statistical decomposition and ML methods used to analyze the high fluctuations in real LV grid data. The first step involves decomposing the time series data using the STL method. This technique decomposes the data into three key components: trend, seasonality, and residual.

- Trend: This component captures the long-term movement in the data, smoothing out short-term fluctuations. It provides a clear view of the overall direction of the grid behavior, which is useful for understanding long-term changes, identifying significant

shifts in data patterns, and providing actionable insights.

- Seasonality: This captures recurring patterns, such as daily or weekly cycles, which can help in identifying predictable changes in the grid, like regular demand peaks. It adjusts for periodic changes, revealing anomalies that might be hidden due to regular fluctuations.
- Residual: The residual component shows the short-term, irregular fluctuations that aren't explained by the trend or seasonality. It highlights sudden or unexpected changes in the grid's behavior, which may indicate anomalies.

Seasonal-trend decomposition is particularly effective for time series data like that from LV grids because it [114]:

1. Manages seasonal fluctuations and sudden changes in trends.
2. Stays robust even when anomalies are present in the data.
3. Handles long seasonal periods effectively.

For detecting anomalies, the focus is on the trend for long-term anomalies and the residual for short-term anomalies. Anomalies in the trend refer to gradual changes in the grid's behavior over time, while anomalies in the residual signal refer to sudden and unexpected shifts.

The model for decomposing a time series, X, is calculated as follows [149, 150]: Consider $X = \{X_1, X_2, \dots, X_n\}$ as a time series data, where each X_i denotes the value at the time i. The focus is on decomposing each data point X_i into its fundamental components: trend, seasonal, and residual, through the use of an additive model. This decomposition is regarded as essential for real-time data analysis, particularly within streaming contexts. The decomposition model is expressed as Eq. (4.36):

$$X_i = T_i + \sum_{p=1}^{P} S_{p,i} + R_i \tag{4.36}$$

Where T_i is the trend component, $S_{p,i}$ represents the seasonal component for p-th ($p = \{1,2,..,P\}$) seasonal period in the series s_p, and R_i is the residual component.

Initially, a convolution filter is applied to effectively isolate the trend component, T_i. The process differs based on the parity of the filter length, f. The filter for an odd filter length is defined in Eq. (4.37), and for an even filter length is defined in Eq. (4.38). The shift by $\left\lfloor \frac{f}{2} \right\rfloor$ is required because the f-point moving average is centered: for odd values of f, the center lies exactly in the middle of the window, and for even values of f, the center lies between two points. Adding $\lfloor f/2 \rfloor$ to the index ensures that this centered trend value is aligned with the correct position in the original time series before detrending.

$$T_i=\frac{1}{f}(X_1 + \ +, \dots, +X_{i+f-1}) \tag{4.37}$$

$$T_i=\frac{1}{f}(0.5 \cdot X_1 + \ +, \dots, +0.5 \cdot X_{i+f-1}) \tag{4.38}$$

The trend component is then subtracted from the original series to isolate variations and facilitate further analysis. The detrend is calculated as Eq. (4.39):

$$DT_i = X_{i+\lfloor f/2 \rfloor} - T_i \tag{4.39}$$

To calculate the seasonality filter, $C_k^{DT}(i)$ is defined as the k-th cyclic subseries of the deterended data set DT. This subseries consists of all data points d_r whose time index r corresponds to the k-th position within a repeating seasonal cycle. Formally, the points included in $C_k^{DT}(i)$ satisfy the condition $r mod s_p = k$, where s_p denotes the seasonality period (for example, 12 for monthly data or 24 for hourly data), and $k \in \{1,2, \dots, s_p\}$. Thus, $C_k^{DT}(i)$contains all detrended observations that fall into the k-th position of each cycle when the entire dataset is grouped into cycles of length s_p.

$$C_k^{DT}(i)=\left\{d_r \mid 1 < r < i \ , \ r \ mod \ , s_p = k\right\} \tag{4.40}$$

This collects all data points from the detrended series DT that correspond to the k -th position in each cycle of length s_p. For the $C_k^{DT}(i) = \left\{d_k, d_{k+ s_p} , d_{k+2 s_p}, \dots\right\}$ the seasonality at period m is as Eq. (4.41). This formula averages or smooths all points in the k -th cyclic subseries up to time i.

$$S_{p,i}=\frac{1}{C_k^{DT}(i)}\sum_{d_j \in C_k^{DT}(i)} d_j \tag{4.41}$$

The detrended series is recalculated for each specific seasonality period s_p, ($s_p = \{s_1, s_2, \dots s_P\}$, $p \in P \mid P = \{1,2, \dots P\}$) as it is explained in Eq. (4.42):

$$DT_{p,i} = DT_{p-1,i} - S_{p,i} \tag{4.42}$$

Finally, the residual component after the trend and seasonal adjustments is computed in Eq. (4.43).

$$R_i = X_{i+\lfloor f/2 \rfloor} - T_i - \sum_{p=1}^{P} S_{p,i} \tag{4.43}$$

Figure 4.13 illustrates the step-by-step decomposition process of a time series X into its trend, seasonal, and residual components, using an additive model. The figure visualizes how raw time series data is progressively processed through filtering and subtraction to isolate distinct structural patterns. This figure highlights the modular and recursive structure of the decomposition method and provides a visual representation of how the components, trend, seasonal, and residual, are systematically isolated from the original time series. (The detrended value DT_iuses the shifted original data point $X_{i+\lfloor f/2 \rfloor}$in accordance with Eq. (4.39).

For simplicity, the index shift is not shown in the figure).

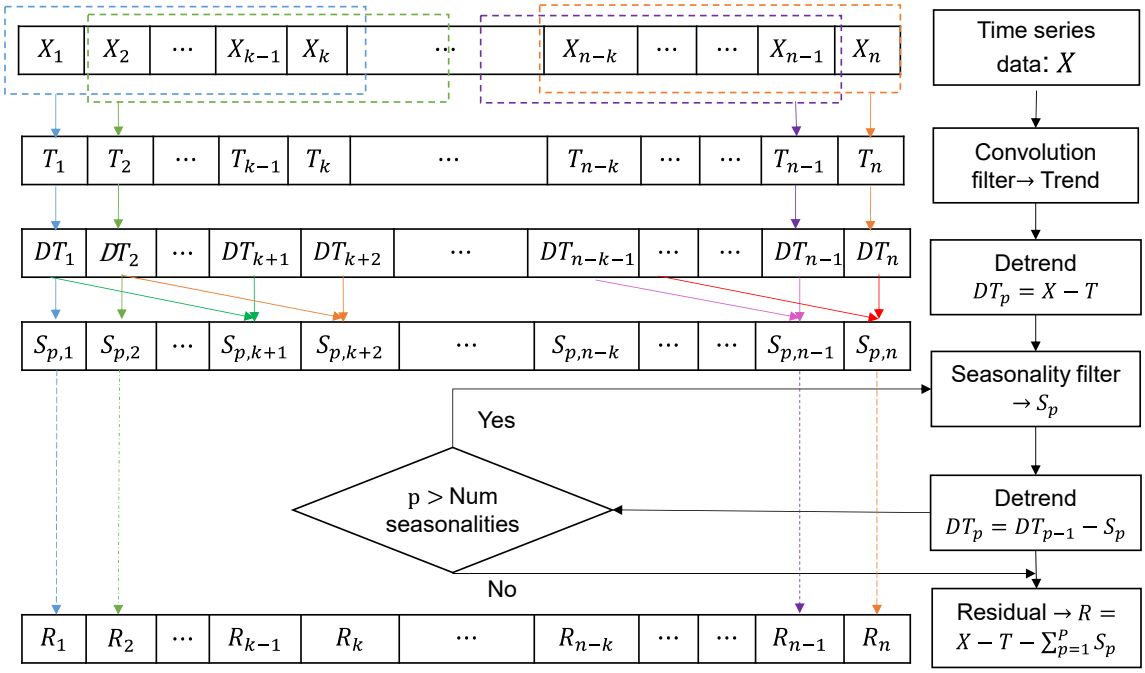


Figure 4.13: Structure of the STL method

Figure 4.14. illustrates the modular workflow for anomaly detection using STL combined with the isolation Forest (iForest) algorithm. The figure highlights the sequential process of decomposing a time series into its core components, trend, seasonal, and residual, and then using these components for targeted anomaly detection. The process begins with the raw time series data, which is first passed through the STL decomposition module. STL separates the input into three distinct components. Each of these components is treated as a distinct input to the isolation Forest model. Isolation Forest is applied to both the trend and residual signals independently to identify context-specific anomalies. Trend anomalies, which represent gradual or sustained deviations in the long-term behavior of the system, and residual anomalies, which indicate sudden, sharp deviations that cannot be explained by expected seasonal or trend patterns. This architecture allows for a dual-layered anomaly detection framework, where each structural aspect of the time series is analyzed separately.

4.3.6 Long Short-Term Memory Autoencoder Method

The LSTM autoencoder is a hybrid approach that combines LSTM networks with an autoencoder architecture to model sequential data for anomaly detection. The proposed LSTM autoencoder acts as a stochastic process model, capturing the normal behavior of the data and allowing the prediction of the probability distribution of the data. Any deviation between the observed data and the predicted probability density distribution is identified as a potential error, fault, or anomaly. LSTM networks are an extension of recurrent NN designed to overcome the limitations of recurrent NN in retaining important information in long sequences.

LSTMs solve this problem by introducing memory cells with feedback connections, enabling the network to keep or discard information over longer periods, selectively [151].

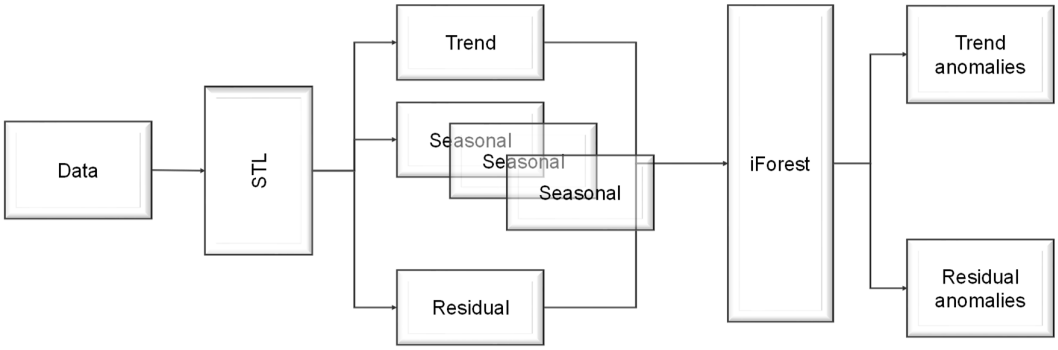


Figure 4.14: Anomaly detection method with the combination of STL and Isolation Forest

As shown in Figure 4.15 each LSTM cell is composed of three key gates: the input gate, which determines what new information should be added to the memory; the forget gate, which decides what information to discard; and the output gate, which controls what part of the memory is used to produce the output. Through these gates, the network can maintain and update the cell's state over time, allowing it to process entire sequences of data effectively.

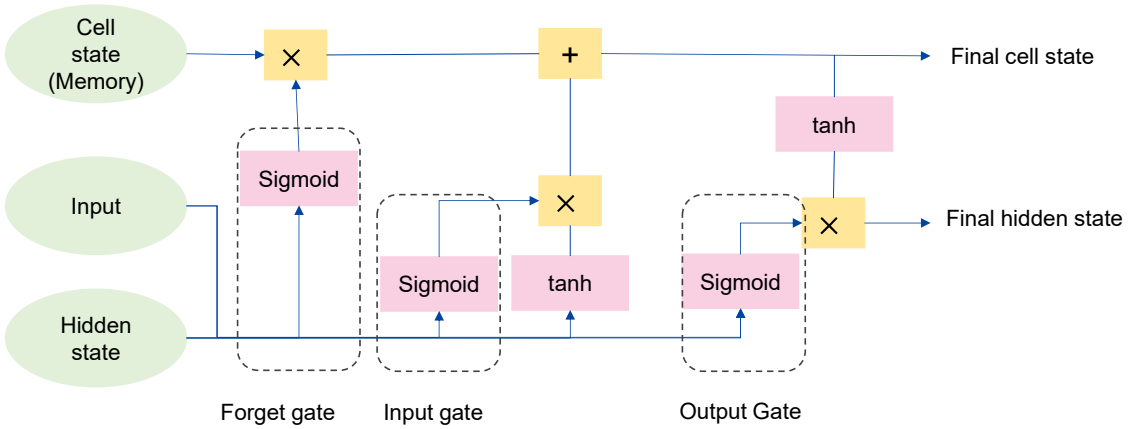


Figure 4.15: Structure of the LSTM method

Autoencoders, on the other hand, are NNs designed to learn efficient representations of data by compressing high-dimensional input into a lower-dimensional space (encoding) and then reconstructing the input (decoding) from this compressed form. The encoder module captures the essential features of the data in a compact representation, while the decoder reconstructs the original data, aiming to minimize reconstruction error. This process also helps reduce noise in the data. Autoencoders are especially popular in anomaly detection tasks because the model is trained to represent normal data patterns.

The LSTM autoencoder first uses its LSTM architecture to learn temporal dependencies within the data and capture short-term and long-term relationships. The encoder then compresses these relations into a dimensionally reduced representation, while the decoder reconstructs

the original sequence from this compressed form. Reconstruction error is used as an indicator of normal or abnormal behavior, where large errors indicate the presence of errors or anomalies in the data.

The combination of LSTM networks and autoencoders is utilized to create a robust method for anomaly detection. LSTM networks are chosen for their ability to capture essential temporal patterns while retaining long-term dependencies, making them ideal for time-series data. On the other hand, autoencoders are employed for their capacity to encode data into a lower-dimensional space and subsequently reconstruct it to capture the most important features. This is particularly useful in scenarios where anomalies are not previously known, allowing the system to identify deviations from the norm based on the learned patterns. By integrating the strengths of both LSTM and autoencoder architectures, the method provides an effective approach for detecting anomalies in complex datasets.

Figure 4.16 shows the structure of the LSTM Autoencoder used in this thesis. The model consists of 4 layers:

1. Input layer: Takes in sequential data, with each sequence representing seven days of energy usage. This sequence length was chosen to capture weekly patterns and potential anomalies while maintaining manageable complexity. A sliding window with a stride of one day ensures that each day's data is included in multiple sequences, providing comprehensive coverage and overlap for effective model training.
2. Structure layer: The structure of the LSTM Autoencoder consists of two LSTM layers followed by a hidden layer in the encoder. The first LSTM layer processes the input sequence, transforming each time step into a 128-dimensional vector to capture temporal dependencies and key features. This output is then passed to the second LSTM layer, which further compresses the 128-dimensional vectors into 64-dimensional representations, distilling the essential features from the data. The final hidden layer further compresses the sequence, potentially utilizing techniques like averaging or attention mechanisms. The decoder part of the autoencoder then reconstructs the original input sequence from this compressed representation.
3. Output layer: The output layer of the LSTM Autoencoder ensures that the decoder's output matches the dimensions of the original input sequence, effectively reconstructing the data from its compressed, encoded state. This ability to reduce the data's dimensionality and then accurately reconstruct it makes the LSTM Autoencoder especially useful for tasks like dimensionality reduction, feature extraction, and anomaly detection in complex datasets.
4. Analysis layer: After training, the LSTM autoencoder is applied to predict test data. Anomalies are detected by calculating the reconstruction loss, which is the difference between the actual data and the reconstructed data from the model. A threshold is set

based on the distribution of these losses, and data points with a reconstruction loss above this threshold are flagged as anomalies. To further analyze these anomalies and understand potential causes, the anomalies are clustered, and the features of each cluster are examined. These features might include factors like temperature, global radiation, or other relevant variables that could explain the anomalies' reason.

4.4 Anomaly Detection for Pattern Change Using Statistical Methods

Statistical methods provide a reliable and interpretable approach for detecting anomalies in power system data. Unlike ML techniques, they require less computational effort and are easier to apply in real-time scenarios. These methods are particularly useful when the data follows known patterns or distributions. In this work, statistical techniques are used to identify unusual behavior in the LV grid, such as faults or unexpected changes in load. The following section introduces and explains the main statistical methods applied in this thesis.

4.4.1 Probabilistic Normalized Expectation Ratio Cumulative Sum

This algorithm aims to detect change points in time series data with a significant variation from the expected pattern. It calculates the ratio of observed data points (y_t) to a Frequent Behavior Weighted Mean (FBWM) (M_t), proposed in this study, which is calculated by clustering historical data (including several years) using the k-means method. Each cluster represents a predominant behavior pattern; the weight assigned to each cluster is proportional to its size, emphasizing more frequently occurring patterns. The weighted mean for each sampling time (t), denoted as $\mathrm{M_t}$, is calculated as follows, Eq.(4.44):

Where $\overline{\mathrm{y}}_{\mathrm{t,k}}$ is the mean of the data points in cluster k at time t, $\mathrm{C_k}$ is the number of observations in cluster k, and k is the total number of clusters.

$$\mathrm{M}_t = \frac{\sum_{k=1}^{K} C_k \cdot \overline{y}_{t,k}}{\sum_{k=1}^{K} C_k} \tag{4.44}$$

This ratio ($\mathrm{R_t} = \frac{\mathrm{y_t}}{\mathrm{M_t}}$) normalizes the data, thereby enhancing the detection sensitivity to deviations from established norms. The probabilistic component of the algorithm employs statistical tests based on the Cumulative Sum (CUSUM) method applied to these normalized ratios, as introduced in [161]. The algorithm presented here is an adaptation of the method described in [162]. The conventional CUSUM algorithm consists of summing the z-standardized values of the time series. A change point is identified when the cumulative total exceeds a predetermined threshold. As a result, this method is classified as an "online" algorithm. The z-standardized is calculated as follows Eq. (4.45):

$$Z_t = \frac{y_t - \overline{y}_t}{\sigma_t} \tag{4.45}$$

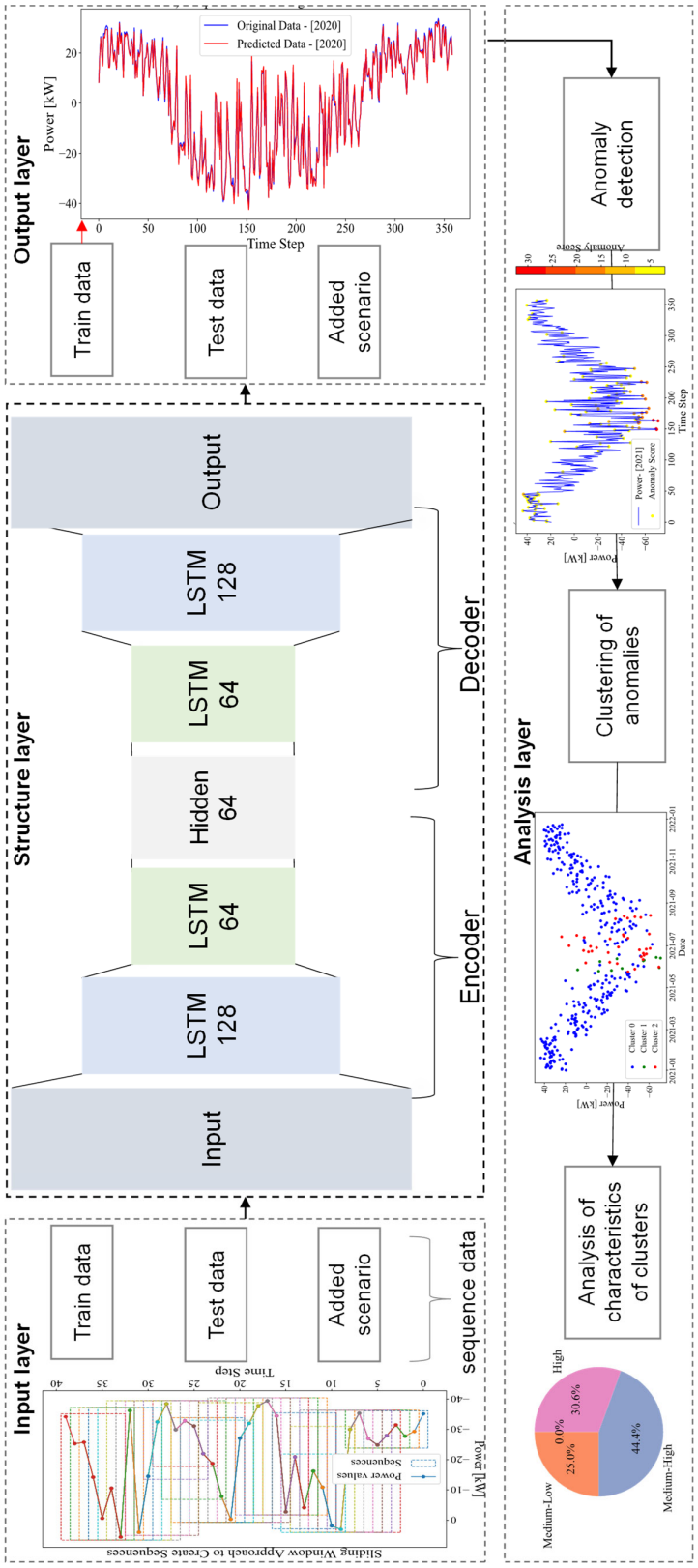


Figure 4.16: Structure of the LSTM Autoencoder

y_t is the observed data point at time t. $\overline{y}_t$ is the mean of the observed data points up to time t. σ_t is the standard deviation of the observed data points up to time t. The algorithm is considered to work by estimating the mean and standard deviation of the series. The standardized sum S_t at time t is calculated as Eq. (4.46):

$$S_T = \sum_{t=1}^{T} Z_t \tag{4.46}$$

The probabilistic version of CUSUM introduces a statistical framework for interpreting the cumulative sum of deviations in the time series data. It takes advantage of the central limit theorem [152] to approximate the distribution of the sum under the null hypothesis, which assumes no change has occurred. When this cumulative sum diverges significantly from expected trajectories, as determined by a statistically derived threshold, a change point is declared.

The central limit theorem in probability theory states that, despite the original variables' distributions, the normalized sample mean tends toward a standard normal distribution under certain conditions. Then it is assumed that by dividing the cumulative sum by the square root of the interval, a standard (theoretical) normal distribution is obtained. Once the warm-up period is completed ($T \geq T_{\text{warmup}}$) (T_{warmup} is the period to estimate initial parameters), the cumulative sum is normalized as Eq. (4.47).

$$\tilde{s}_T = \frac{S_T}{\sqrt{T}} \tag{4.47}$$

Probability of change at time t, $Prob_t$, is calculated using the standard normal cumulative distribution function; this probability is used instead of the raw standardized CUSUM sum for change point detection. Equations (4.48) and (4.49) are proved in the Appendix A.

$$\Phi(\tilde{s}_t) \approx P\left(\widetilde{S_T} \leq \tilde{s}_t\right) \tag{4.48}$$

$$Prob_t = 2 \cdot (1 - \Phi(|\tilde{s}_t|)) \tag{4.49}$$

Where $\Phi(.)$ is the standard normal cumulative distribution function (CDF). During the warm-up phase ($1 \leq t \leq T_{\text{warmup}}$), the cumulative sum is updated solely to estimate the mean and variance required for standardization, and no change detection is performed. At $t = T_{\text{warmup}}$, the algorithm transitions from parameter estimation to monitoring. For $t > T_{\text{warmup}}$, the normalized CUSUM statistic is used to compute the probability of change Prob_t. A change point is detected if the probability of change exceeds a predefined threshold ($\text{Prob}_t > p_{\text{limit}}$.). The parameter p_{limit} controls the sensitivity of the change detection method. The flowchart of this algorithm is presented in Figure 4.17.

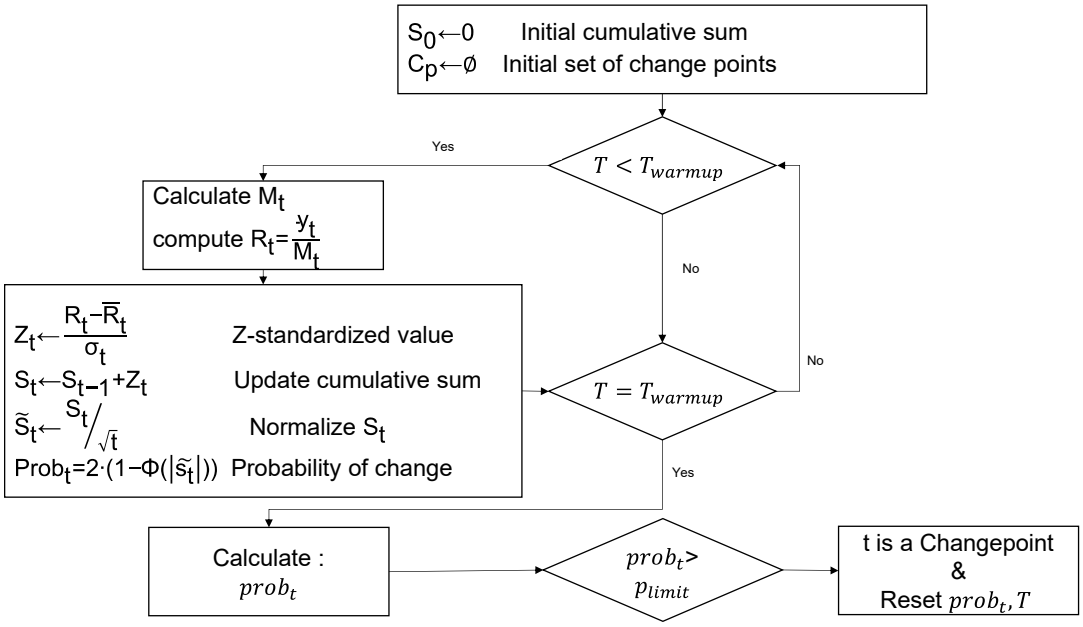


Figure 4.17: The flowchart of the CUSUM method

4.4.2 Seasonality Analysis Using Probability Density Function Methods

This section focuses on the identification of weekly seasonal patterns in time series data and their comparison across multiple years in order to detect anomalies. The seasonal patterns are highly consistent across multiple years, and any deviations from these patterns can indicate potential anomalies.

Time series data in LV grids exhibit strong weekly patterns due to routine activities on specific days of the week. To compare these patterns accurately across different years, it is crucial to align the data so that each series starts on the same weekday. By aligning each time series to start on the same weekday, the precision of analysis is enhanced, and this reduces noise and enhances the clarity of the pattern comparisons. The alignment process includes:

- Identify the start day: Determine the starting day of each time series.
- Realign data: Shift the data points so that all series start on the same weekday.
- Adjusting: Modify the length of the time series to ensure consistency across the dataset.

Kernel Density Estimation (KDE) analyzes the distribution of the aligned time series data. This non-parametric technique estimates the probability density function (PDF) of a random variable, providing insights into the smoothed data distribution that helps reveal underlying patterns [163]. The KDE for a set of data points $\{y_i\}_{i=1}^{n}$ is calculated as follows Eq. (4.50):

$$\hat{f}(y) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right) \tag{4.50}$$

Where $\hat{f}(y)$ is the estimated density at point y, n is the number of data points, $\mathrm{K}(.)$ is the kernel function (typically a Gaussian distribution), h is the bandwidth, a parameter that controls the smoothness of the resulting density curve, which is selected in this thesis based on Scott's rule. The values y_i represent the observed data points.

The choice of the kernel $\mathrm{K}(.)$ and the bandwidth h significantly affects the shape of the estimated PDF. The Gaussian kernel is a common choice due to its smoothness properties in Eq. (4.51):

$$K(y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}} \tag{4.51}$$

The KDE provides a smooth estimate of the data distribution, highlighting the areas of high and low density. By comparing the estimated PDFs for different periods, deviations from the expected pattern can be identified, indicating potential changes in the underlying process.

To quantify deviations between weekly distributions, the Kullback–Leibler Divergence (KLD) is applied. KLD measures the divergence between two probability distributions and is used here to compare the KDE of a given week with a reference distribution representing normal system behavior [164]. The KLD between two distributions P and Q is defined as Eq. (4.52).

$$KLD(P \parallel Q) = \sum_{y \in Y} p(y) \log \left(\frac{p(y)}{q(y)}\right) \tag{4.52}$$

Where $P$and Q denote the probability distributions obtained from KDE. By evaluating the KLD between corresponding weekly distributions across different years, changes in seasonal behavior can be systematically identified.

4.4.3 Dynamic Threshold Clustering Method

To effectively detect anomalies in energy consumption, the dynamic threshold clustering method leverages scenario-based clustering and adaptive boundary estimation to account for the variability of contextual conditions such as weather and time. This method dynamically adjusts detection criteria based on the specific conditions under which energy usage occurs. By segmenting the data into quantile-based clusters that reflect distinct operational scenarios, defined by features such as temperature, solar radiation, time of day, and day type, the method ensures that comparisons are made within contextually relevant groups.

Data features such as temperature, global solar radiation, hours of the day, and day type (workdays versus weekends) are categorized into distinct scenarios as follows:

- Temperature(T): $T \in \{Low, Medium\ Low, Medium\ High, High\}$
- Global solar radiation (G): $\mathrm{G} \in \{\mathrm{Low}, \mathrm{Medium\ Low}, \mathrm{Medium\ High}, \mathrm{High}\}$
- Hours of the day (H): $H \in \{\mathrm{Midnight}, \mathrm{Morning}, \mathrm{Afternoon}, \mathrm{Evening}\}$
- Day type (D): $D \in \{\mathrm{Workday}, \mathrm{Weekend}\}$

Clusters are defined by the Cartesian product of these categories, resulting in a comprehensive set of potential scenarios that the power grid may encounter, encapsulating 164 unique combinations Eq. (4.53).

$$C = T \times G \times H \times D$$
$$| C | = 4 \times 4 \times 4 \times 2 = 128 \tag{4.53}$$

Within each cluster $c \in C$, historical power measurements are analyzed to determine the minimum and maximum observed power values, denoted by $P_{\min}(c)$ and $P_{\max}(c)$. These values define the baseline operating boundaries under the corresponding conditions and serve as a reference for anomaly detection.

To ensure the robustness of these boundaries, historical outliers are removed prior to boundary estimation. Outlier detection is performed using the interquartile range (IQR) method. A power value is classified as an outlier if it lies outside the interval: $[Q_1 - k \cdot IQR\,,\ Q_3 + k \cdot IQR]$.

where $Q_1$and Q_3denote the 25th and 75th percentiles of the power data, respectively, $\mathrm{IQR} = Q_3 - Q_1$, and $k$is a sensitivity coefficient, set to $k = 1.5$ in this study (the sensitivity parameter k is chosen as $k = 1.5$, in accordance with the standard IQR-based outlier detection approach proposed by Tukey [153]). Tukey chose the value 1.5 because it extends the middle 50% of the data to include normal fluctuations while also clearly identifying unusual values. After removing outliers, the power boundaries $P_{\min}(c)$ and $P_{\max}(c)$are recalculated for each cluster. These refined boundaries provide a reliable baseline for evaluating new observations.

For a new data point observed at time t, characterized by the feature vector (T_t, G_t, H_t, D_t), the corresponding cluster $c \in C$ is identified. The measured power value P_t is then compared against the cluster-specific boundaries. According to Eq. (4.54), power values outside the cluster-specific boundaries indicate deviations from normal operating behavior.

$$Anomaly = \begin{cases} P_t < P_{min}(c) \\ P_t > P_{max}(c) \end{cases} \tag{4.54}$$

To find the score of anomalies and the data that are too close to the maximum and minimum border, the error for the minimum and maximum border are calculated as Eq. (4.55):

$$E_{max} = P_t - P_{max}(c)$$
$$E_{min} = P_t - P_{min}(c) \tag{4.55}$$

Based on these equations, the anomaly score is calculated as the absolute deviation of the measured power from the corresponding cluster-specific boundary. These deviation values are used to assess how close a data point is to the minimum or maximum power limits. Data points whose deviations lie below the anomaly boundary but within a predefined margin are

classified as alarm points, indicating operating conditions that are close to becoming an anomaly.

In this study, a threshold of 5 kW is applied to distinguish alarm points from normal operating conditions. The value of 5 kW is chosen as a conservative margin based on the observed variability of power measurements, ensuring that minor fluctuations are not misclassified as alarm conditions.

4.5 Anomaly Detection for Identifying Unregistered Installed PV Systems

In modern distribution grids, the integration of rooftop PV systems has significantly increased, but not all installations are promptly registered or reported to the grid operator. This lack of visibility poses a major challenge for Distribution System Operators (DSOs), as unregistered or delayed registrations of PV systems can result in incorrect assessments of local generation capacity, affect grid stability, and hinder the planning of flexibility resources. The root of the problem lies in the limited digitalization of reporting systems, as well as the decentralized and often informal nature of small-scale PV installations. In many cases, operators do not have timely access to information about how many panels were installed, their capacity, exact location, or real-time generation data. Sometimes, only the installed capacity is known, if at all, without interval-based monitoring or accurate forecasting, creating a “black box” situation in parts of the grid.

To address this gap, this method introduces a data-driven approach for detecting potentially unregistered PV systems using anomaly detection. It focuses on situations where only net power measurements, which are the sum of consumption power (positive values) and PV generation output power (negative values), are expressed as $P = P_{con} + P_{pv}$. The main purpose of this section is to separate the PV generation data from the combined power readings and detect unregistered PV in the grid.

In the scenario-based approach, PV output was calculated leveraging global radiation data recorded at 10-minute intervals. This involved employing various scenarios corresponding to different PV panel types, namely 60-cell, 72-cell, and 96-cell panels. The formula used to determine the power output of the PV panels is Eq. (4.56):

$$P_{outPV} = \eta_{PV} GS(t) \cdot A_{panel} \tag{4.56}$$

Where $GS(t)$ represents the solar irradiance at time t, and A_{panel} is the total area covered by the solar panels, and η_{PV} represents the overall conversion efficiency of the PV system. The efficiency η_{PV} is assumed to be equal to 1, representing an idealized maximum power scenario. This conservative assumption ensures that the estimated PV output corresponds to an upper bound, thereby increasing the sensitivity of the anomaly detection method to the presence of unregistered PV systems. The number of panels, N_{panel}, is calculated based on the installed PV capacity $P_{installed-PV}$ and the panel capacity P_{panel}, which is presented in Eq. (4.57). In this

study, typical nominal panel capacities of 250 W and 400 W are considered, corresponding to commonly deployed residential PV modules.

$$N_{panel} = \frac{P_{installed-PV}}{P_{panel}} \tag{4.57}$$

To estimate the total area of the solar panels, the following scenarios are considered:

- 60-cell solar panel: 39” W x 66” L=1.629 m^2
- 72-cell solar panel: 39” W x 77” L=1.936 m^2
- 96-cell solar panel: 41.5” W x 62.6” L=1.676 m^2

Thus, the total area covered by the solar panels is calculated using Eq. (4.58):

$$A_{panel} = N_{panel} \cdot S_{Panel}$$

$$S_{Panel} = \begin{cases} 1.659\ \text{m}^2 & \text{if } 60 \text{ Cell} \\ 1.936\ \text{m}^2 & \text{if } 72 \text{ Cell} \\ 1.676\ \text{m}^2 & \text{if } 96 \text{ Cell} \end{cases} \tag{4.58}$$

For analysis, the scenario yielding the highest PV output is selected to ensure that the calculated P_{PV} represents the maximum probable contribution of the PV system. This is essential to accurately gauge the potential reduction in actual consumption due to solar power generation, especially during peak irradiance periods. Once the maximum P_{PV} is estimated, it is deducted from the total power recorded (P) to isolate the net power consumption ($\mathrm{P_{con}}$). This is expressed by Eq. (4.59):

$$P_{con} = P - P_{pv} \tag{4.59}$$

With the normalized net power data (P), the K-means clustering algorithm is applied daily to classify days into distinct groups according to their consumption patterns. For this analysis, four clusters are selected to capture representative variations in daily energy usage behaviors.

The results are expected to show mostly positive values, assuming that no energy storage devices or electric vehicle discharge scenarios affect the grid. This reflects days that are only characterized by consumption patterns. However, if the consumption patterns yield negative values, especially in clusters where the net power is minimal, indicating low consumption and high PV production, there may be unregistered PV systems.

4.6 Anomaly-Based Detection of Outages

This subchapter addresses a critical challenge in LV grid management: the limited visibility caused by the lack of metering at all nodes. In most LV grids, especially older or less digitized ones, real-time monitoring is only available at a few key locations, typically on the secondary side of transformers or at busbars, leaving the majority of the grid unobserved. This lack of granularity makes it difficult to detect localized faults or outages in a timely and accurate manner, particularly when the affected nodes are not directly monitored. To overcome this gap,

an anomaly-based detection approach is proposed. By analyzing patterns in the power data collected from transformer-connected monitoring points, the method enables the identification of unusual behaviors that may indicate outages at downstream nodes.

The anomaly detection analysis begins with a data normalization step that differs from the normalization approach used in the previous anomaly detection framework. This normalization is specifically designed to emphasize deviations in grid parameters caused by outage events, thereby improving the separability between normal and abnormal operating conditions. The isolation Forest algorithm is then applied to the normalized data, focusing solely on data from buses connected to transformers to assess whether these critical points can effectively indicate outages. To further enhance the accuracy and reliability of outage detection, the methodology incorporates both active (P) and reactive (Q) power measurements from the grid.

The data is prepared by computing the differences between consecutive measurements of active and reactive power, as shown in the following Eq. (4.60):

$$\Delta P_t = P_t - P_{t-1} \tag{4.60}$$

This difference (ΔP_t) highlights sudden changes in power values that could indicate disruptions such as outages. After data preparation, isolation Forest models are trained for ΔP. The training data sets represent normal operating conditions, without outages. This training helps the models to establish baselines for detecting deviations that are characteristic of outages.

The current anomaly-based methods for outage detection in LV grids exhibit several limitations that are closely related to the radial topology of these networks and the heterogeneous distribution of loads. In radially operated LV grids, outages occurring at nodes located far from the transformer or at nodes with low demand may not produce pronounced variations in active power measurements at transformer-connected buses. As a result, such outages can remain difficult to detect using power-based indicators alone.

To overcome these challenges, the proposed method incorporates reactive power (Q) analysis into the anomaly detection process. Reactive power is generally less sensitive to changes in load types that are common in LV grids. The anomalies are detected for ΔQ ($\Delta Q_t = Q_t - Q_{t-1}$). The analysis employs a decision logic to enhance detection accuracy: when an anomaly is detected in active power (ΔP) but not in reactive power (ΔQ), the observed deviation is more likely attributable to normal load variations rather than to an actual outage.

To formalize this distinction, a conditional probability framework is applied. Let O_P denote the event that an anomaly is detected in active power, and let N_Q denote the event that reactive power remains within normal operating conditions. The probability of an outage indicated by active power, given normal reactive power behavior, is computed as follows, Eq. (4.61). Figure 4.18 illustrates the flowchart of the outage detection process.

$$P(O_P|N_Q) = \frac{P(O_P \cap N_Q)}{P(N_Q)} \tag{4.61}$$

This approach focuses on isolating true outages in active power by leveraging reactive power stability to reduce false positives. To improve the sensitivity of outage detection under these conditions, additional electrical measurements are required. In particular, current transformers capable of providing negative- and zero-sequence current components can offer valuable information, as these quantities are highly sensitive to network asymmetries and unbalanced operating conditions that commonly arise during fault or outage events. The inclusion of sequence-based measurements, therefore, enhances the detectability of localized outages that may not be observable through power-based indicators alone.

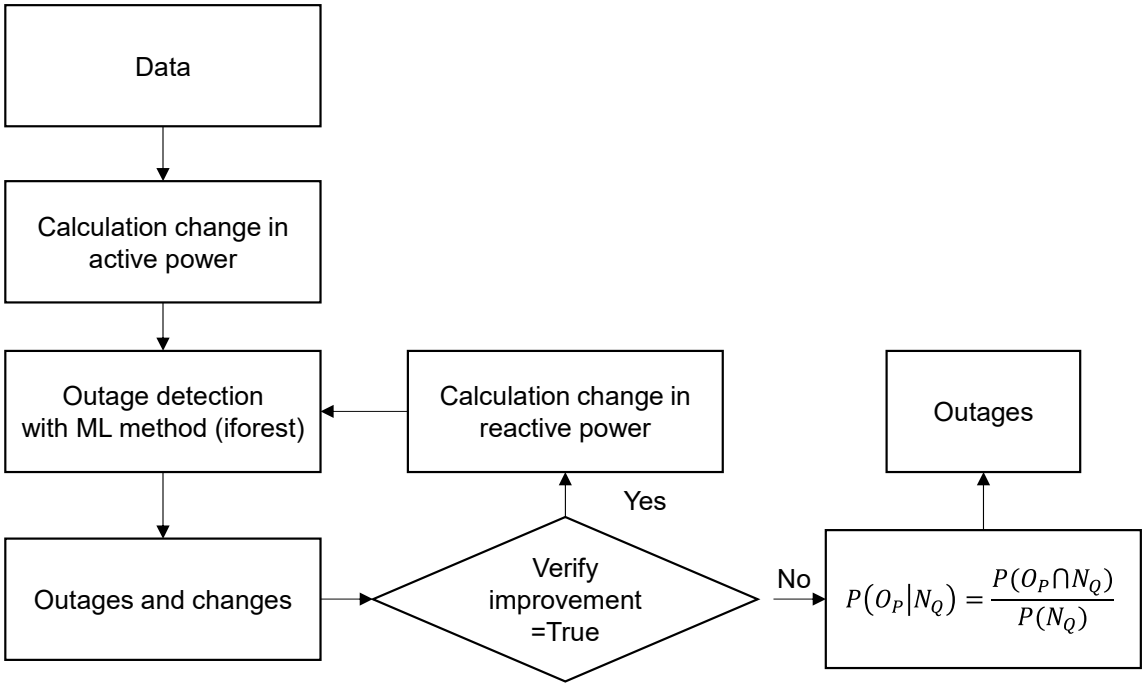


Figure 4.18: Flowchart of outage detection

4.7 Summary and Research Focus

This chapter presented a data-driven methodology aimed at improving monitoring and situational awareness in LV distribution grids through two main approaches: PWCF and anomaly detection. These methods were selected to address the increasing complexity and uncertainty in modern LV grids caused by integrating renewable energy sources, electric vehicles, and new loads. Traditional forecasting techniques often fail in this context due to long sampling intervals, model inaccuracies, and the stochastic behavior of prosumers

The PWCF method was chosen as it provides upper and lower boundaries rather than point forecasts, allowing grid operators to anticipate and mitigate voltage or current violations. It is particularly suited for preventive control, where the goal is to ensure safe operating margins without requiring exact predictions. This method was further enhanced using NNs to increase

its adaptability to real data fluctuations, leveraging historical data to adjust forecast bounds dynamically.

The anomaly detection framework incorporates both ML and statistical models to identify irregularities in energy consumption, unregistered PV systems, and operational faults. These models were selected for their ability to handle noisy, high-variance datasets and operate without full visibility of the grid topology. Approaches include unsupervised techniques such as LSTM autoencoders and isolation Forests, which are well-suited for detecting unknown or rare anomalies in sequential data.

Most LV grids suffer from low-resolution measurements, irregular sampling, and incomplete labeling of events, making it difficult to train supervised models reliably. Additionally, small-scale anomalies often remain hidden in training data, increasing the risk of missed detections or false positives. The reliance on external data, such as weather or demand forecasts, adds further uncertainty when those sources are imprecise or poorly synchronized. These limitations underline the importance of using robust, data-driven methods like PWCF and unsupervised anomaly detection, which can adapt to sparse, incomplete, or noisy datasets without requiring fully labeled data or detailed system models. Future work should focus on improving data quality, exploring self-supervised learning, and enhancing the interpretability of these models for practical deployment.

5 Case Studies and Simulation Results

This chapter presents the practical evaluation of the proposed forecasting and anomaly detection methods using multiple real and synthetic datasets. The aim is to validate the effectiveness and robustness of the developed technique, namely the Pseudo-Worst-Case Forecasting (PWCF), Machine Learning (ML)-based anomaly detection, and statistical anomaly detection.

The chapter is organized into two main parts: First, Section 5.1 presents a detailed overview of the datasets employed in the analysis, explaining their sources, structure, and relevance to the case studies. Then, Section 5.2 provides simulation results and evaluations corresponding to each of the core methods proposed in Chapter 4.

5.1 Introduction to Case Studies and Datasets

Throughout this thesis, data from three separate case studies have been considered. First, smart meter data from the SmartAPO[8] project, hereafter referred to as "SMAPO," will be analyzed. Second, the net power data from the intelligent Ortsnetzstationen (iONS) (local network substations), which are further explained in the Appendix B, and linked to the AISOP[9] project, are examined. Finally, the third case study includes an LV grid for Simbench open-source data.

5.1.1 Measured Voltage and Current Data from Smart Meters (SmartAPO Project)

The SMAPO data originates from a measurement series conducted over one year (2015) within an LV grid in Germany. This particular grid primarily supplies residential households and does not include industrial customers. The measurements were part of the SmartAPO project by the Technical University of Kaiserslautern. Measurements were taken at 193 nodes, recording voltages and phase angle available across all three conductors (L_1, L_2, L_3). Data was collected at 10-minute intervals, resulting in 52,560 voltage readings per node for the year. Current measurements were also taken in three phases at the same interval, but at only 102 measurement points. The grid included 23 Photovoltaic (PV) installations, which feed into the grid at various nodes. (Details are presented in the Appendix C)

5.1.2 Real Net Power Data (AISOP Project)

The methodologies were applied using real data obtained from WWNetz, a distribution system operator (DSO) in Germany. The data set, received from over 100 LV stations, includes a directional net power time series recorded every 15 minutes from the secondary side of LV transformers across various grids, installed PV capacity, and installed transformer capacity.

[8] Smart Autonomous Predictive Operating
[9] AI-assisted grid situational awareness and operational planning

Notably, the dataset lacks information such as the number and characteristics of PV systems at each station, their precise geographic locations, and specific energy consumption data.

Information on whether stations are in urban or rural areas, crucial for regional analysis, is unavailable. Supplementary environmental data, including temperature, wind speed, global solar radiation, and sunshine hours, were obtained from the Deutscher Wetterdienst (DWD) [154]. To address data gaps and improve the analysis, the study selectively focused on stations that exhibit high fluctuation behavior in net power. These stations were chosen based on an algorithm described in Appendix E. To validate the anomaly detection algorithms across different station types, a classification algorithm developed in this thesis was used to identify whether each station is more likely to be urban or rural, as described in the Appendix D.

For simulation purposes, two representative stations were selected, as shown in Table 5.1:

- The urban station is equipped with a 630 kVAR transformer and PV installations that were upgraded in 2015 from 22 kW to 41 kW. The ratio of PV to transformer capacity at this station is 0.065.
- In contrast, the rural station features a 400 kVAR transformer and 215 kW of installed PV capacity, resulting in a significantly higher capacity ratio of 0.53. However, data from the years 2015 and 2012 are missing for this station, highlighting gaps in historical data continuity.

Furthermore, to evaluate the algorithm using real-time data, several actual stations within WWNetz were selected for online monitoring. These stations stream live three-phase voltage and current data, enabling the testing and validation of proposed algorithms on an advanced digital platform. To facilitate this, a Long Range Wide-Area Network (LoRaWAN) sensor has been employed. LoRaWAN, part of the Low Power Wide Area Network (LPWAN) family, specializes in long-range, low-power communication, making it ideal for remotely collecting and transmitting data across vast distances [155].

Table 5.1: Information on two selected LV stations

Station		**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Station 1 (Urban)	Transformer capacity [kVA]	630									
	Installed PV [kW]	22	22	22	41	41	41	41	41	41	41
Station 2 (Rural)	Transformer capacity [kVA]	400									
	Installed PV [kW]	-	215	215	-	215	215	215	215	215	215

5.1.3 SimBench-Based Data

SimBench is a research project that ran from 2015 to 2019 as part of a German government energy program. It was conducted in close collaboration between the ie³ Institute at TU Dortmund (Prof. Rehtanz), the University of Kassel, and the IAEW at RWTH Aachen. The project created a database that includes information about electrical loads and the layout of electrical grids, focusing on LV grids [156, 157].

The LV grid scenarios in SimBench cover different setups, from rural to urban environments, and include detailed information about the types of loads and the grid structure. This makes it useful for simulations to analyze how electrical grids perform under various conditions. The SimBench data is openly available to researchers and engineers for academic and practical applications.

This thesis selects the urban LV grid scenario named "1-LV-urban6--1-sw" for detailed investigation. This naming convention in the SimBench dataset encodes several important characteristics: '1' indicates the SimBench version, 'LV' specifies the voltage level as' Low Voltage', 'urban 6' represents the sixth urban configuration, which is characteristic of the urbanization of the region. The double dash '--1' represents a near future scenario ('0' = current state, '1' = near future, '2' = long-term future). The suffix 'sw' indicates the inclusion of a switching cabinet in the model. These components define the framework of the selected urban LV scenario, whose key characteristics are presented in Table 5.2.

Table 5.2: Key features of LV-urban6 [158]

Topology	Transformer [kVA]	Number of feeders	Overall length [km]	Number of loads	Overall load [kW]	Number of DER[10]	Installed DER power [kW]
Radial Network	630	7	1.08	111	441	5	58

5.2 Simulation Results

This chapter presents the results of the methodologies introduced in Chapter 4, applied to the use cases. As shown in Figure 5.1, the developed methods generate decision-support signals for operators, including alarm and emergency signals, which assist in event control and decision adjustments. Additionally, evaluated data, comprising both analyzed and forecasted data, are provided either periodically or upon request, in line with the flexible data exchange concept of the Digital Process Twin (DPT).

[10] Distributed Energy Resources

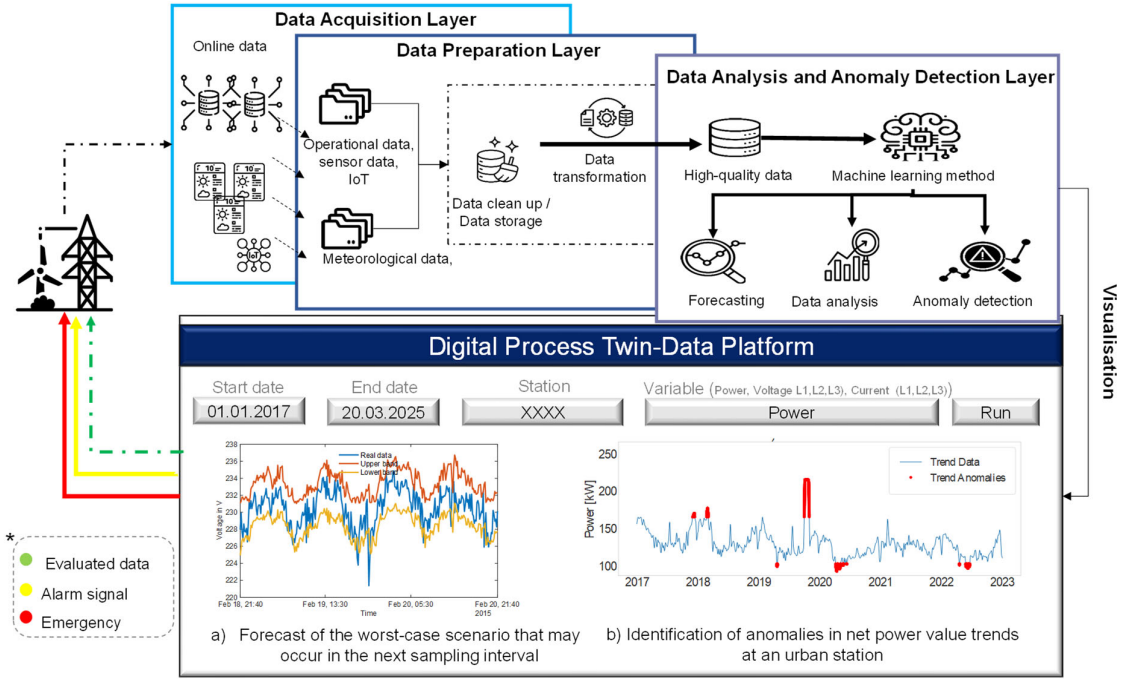


Figure 5.1: Architecture of the DPT enabling forecasting, analysis, and anomaly detection

5.2.1 Results of Methods: Pseudo-Worst Case Forecast and Pseudo-Worst Case Forecast with Neural Network

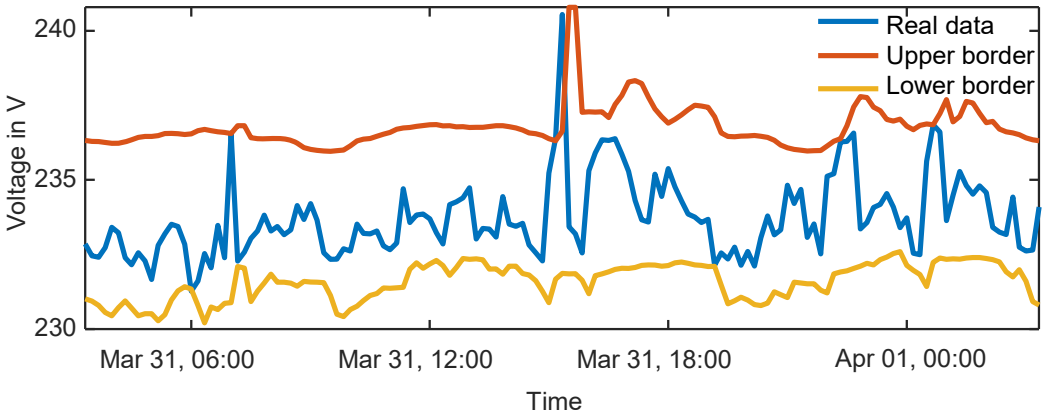
The PWCF and PWCF with Neural Network (NN) (PWCF + NN) methods were simulated using the SMAPO dataset to assess their performance. In these simulations, complete state information is available for each time step and is used as input for the forecasting models. It is important to note, however, that in real-world applications, this state information would typically be affected by measurement and estimation errors.

In both simulations, the parameter “D=28" indicates that data from the previous 28 days are considered in the analysis, providing historical context for forecasting.

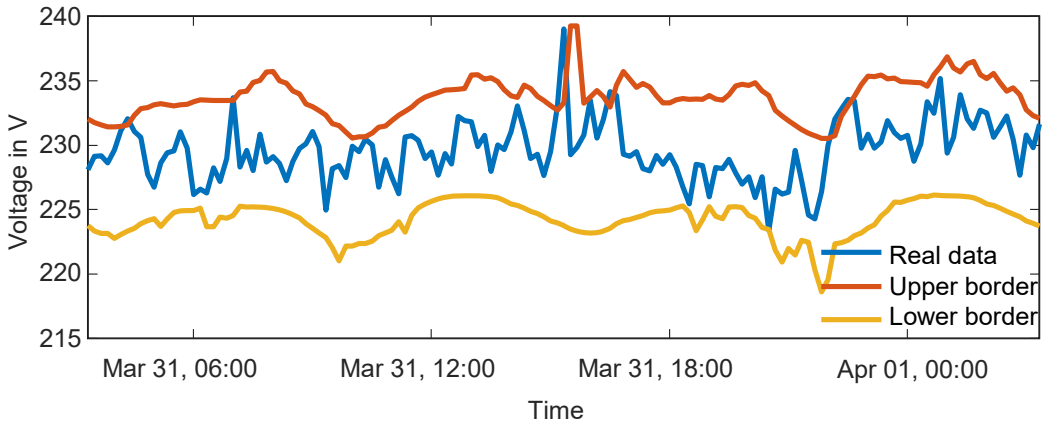
To further evaluate method performance, results for the absolute values of voltages and currents are analyzed. The SMAPO dataset contains real LV measurement data that exhibit substantial variability, primarily due to voltage adjustments performed by the tap changers of the upstream transformer, which acts as the slack node. These tap-change operations introduce visible step-like variations in the LV voltages, making short-term forecasting more difficult. This variability significantly complicates the task of forecasting sudden and substantial changes, which often lead to the largest forecasting absolute errors during the most challenging days, referred to as 'worst days'. Despite these challenges, the PWCF method proves highly effective for most other periods.

Figure 5.2 illustrates the outcomes of the PWCF method applied to voltage values for the upper border. Specifically, Figure 5.2 (a) displays the highest voltage value observed among all nodes throughout the year. Figure 5.2 (b) details the maximum absolute error in the PWCF

method for forecasting the upper border, which occurs during high fluctuations. This represents the highest error for the upper border and often happens on the worst days.



a) the highest voltage value throughout the year (PWCF method)



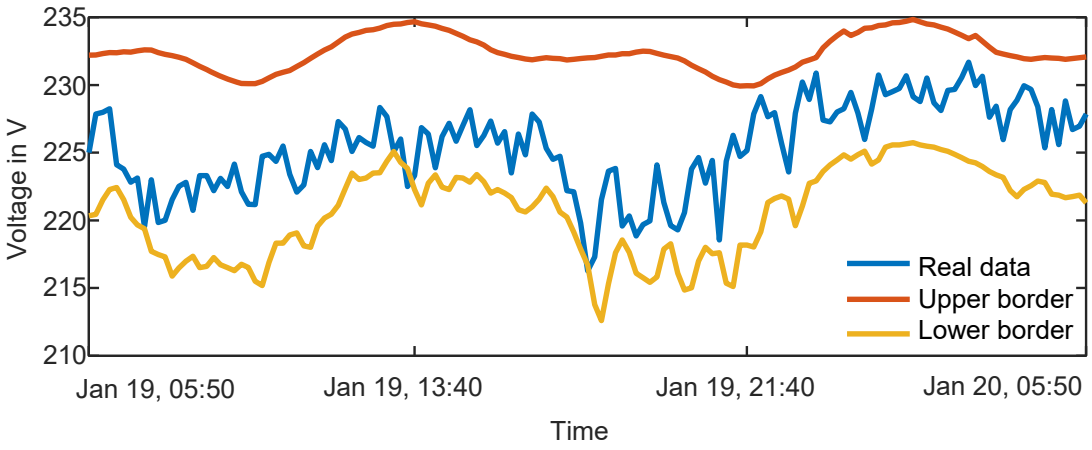
b) the highest error in the upper voltage border (PWCF method)

Figure 5.2: Results of the PWCF method applied to voltage data for upper borders. (a) highest yearly voltage across nodes, (b) maximum error in upper border forecasting

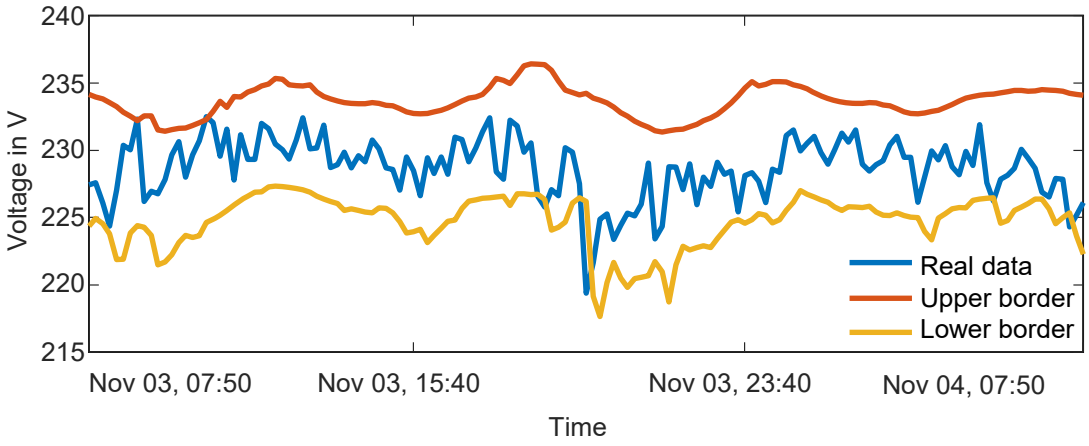
Figure 5.3 presents the results for the lower border. Figure 5.3 (a) shows the lowest voltage value recorded across all nodes throughout the year, while Figure 5.3 (b) highlights the maximum forecasting error for the lower boundary, which typically occurs during periods of significant fluctuation and represents the worst-case scenario. This analysis emphasizes the days characterized by the most extreme errors, highlighting the robustness of the forecasting method under severe testing conditions, including the highest and lowest absolute values of voltages and currents recorded throughout the year.

Figure 5.4 (a) highlights the days with the highest current magnitudes over the year. Figure 5.4 (b) illustrates the maximum error in forecasting the upper border for currents on the worst day, emphasizing the challenges posed by sudden and significant changes, such as rapid load shifts or the impact of control on current flow (Although the LV grid itself does not include active control, occasional tap-changer operations at the upstream MV/LV transformer can indirectly

influence the measured current and voltage profiles observed by smart meters). The maximum recorded error in some instances reaches 80 A. Although occasional inaccuracies occur during extreme fluctuations, the results demonstrate that such current spikes typically exceed the forecasted value only briefly. The forecasting method maintains a high level of accuracy, even under conditions of substantial current variability.



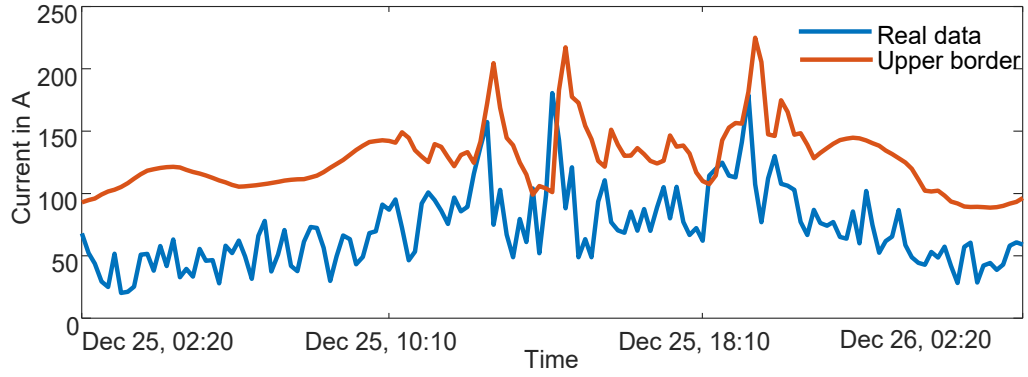
a) the lowest voltage value throughout the year (PWCF method)



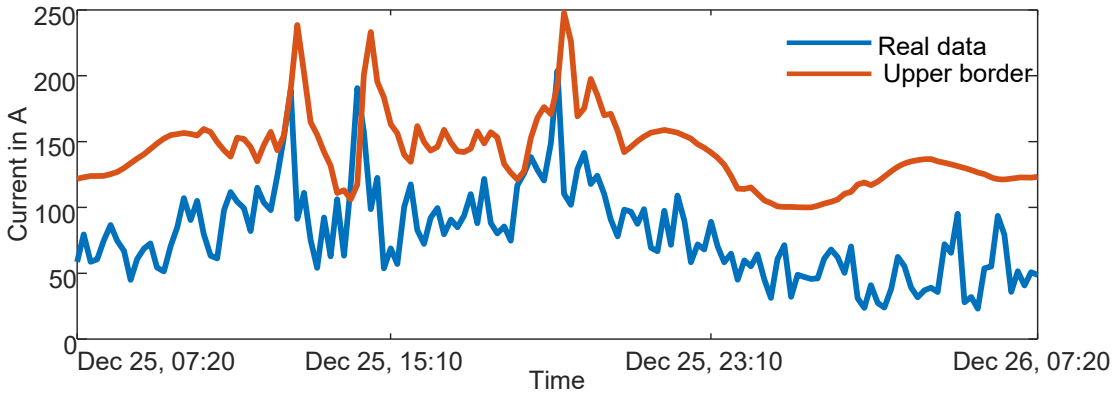
b) the highest error in the lower voltage border (PWCF method)

Figure 5.3: Results of the PWCF method applied to voltage data for lower borders. (a) lowest yearly voltage across nodes, (b) maximum error in lower border forecasting

The success rates (Success rate (%)=(1-$\frac{\text{Number of error}}{\text{Total samples}}$) ×100) of the PWCF method applied to the SMAPO dataset are as follows: 98.388% for the upper voltage border, 98.216% for the lower voltage border, and 98.167% for the current border. As detailed in section 4.2.1, the effectiveness of this method strongly depends on the operator's knowledge and expertise regarding the grid's behavioral patterns. This dependency arises from the need to manually select the target variables and define appropriate upper and lower boundaries. To reduce this reliance and improve forecasting accuracy, an enhanced version of the method, the PWCF method, is combined with NN, which is discussed in section 4.2.2.



a) the highest current value throughout the year (PWCF method)



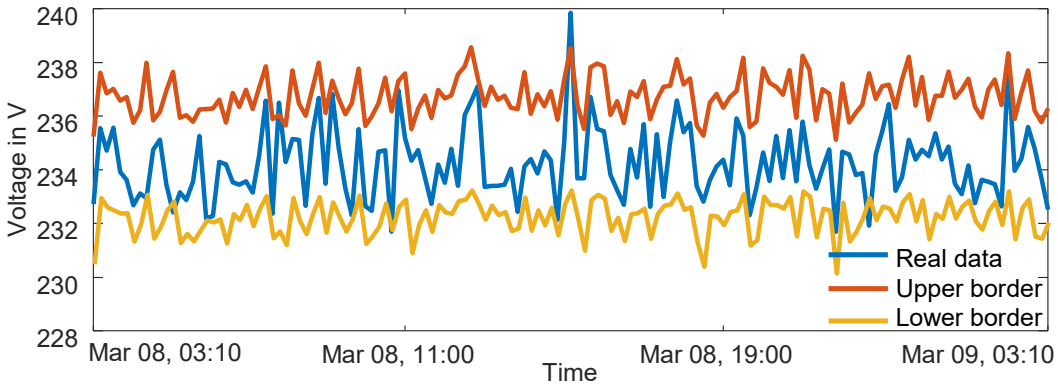
b) the highest error in the upper current border (PWCF method)

Figure 5.4: Results of the PWCF method applied to current data (a) highest yearly current across nodes, (b) maximum error in upper border forecasting

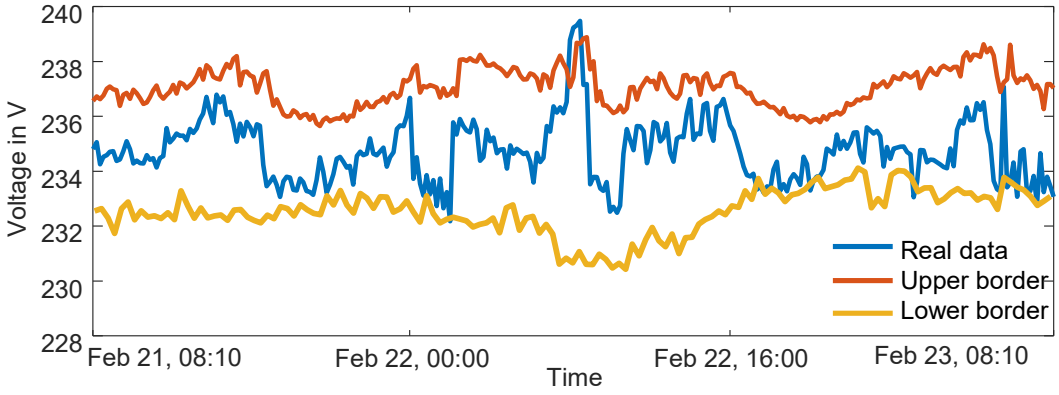
Figure 5.5 illustrates the outcomes of the PWCF+NN method for the upper voltage boundary. Figure 5.5 (a) displays the highest voltage value observed across all nodes during the year, while Figure 5.5 (b) shows the maximum forecasting error for the upper boundary on the worst day, reflecting the challenges of sudden voltage spikes.

Figure 5.6 presents the corresponding results for the lower voltage boundary. Figure 5.6 (a) highlights the lowest voltage value recorded across all nodes, and Figure 5.6 (b) depicts the maximum forecasting error for the lower boundary during the worst-case scenario, emphasizing the difficulty in capturing sudden voltage drops.

Similarly, Figure 5.7 provides the results related to current magnitudes. Figure 5.7 (a) shows the days with the highest current magnitudes observed during the year. Figure 5.7 (b) illustrates the maximum error in forecasting the upper boundary for current on the most challenging day, where rapid load variations and the effects of control commands are apparent.



a) the highest voltage value throughout the year (PWCF+NN method)

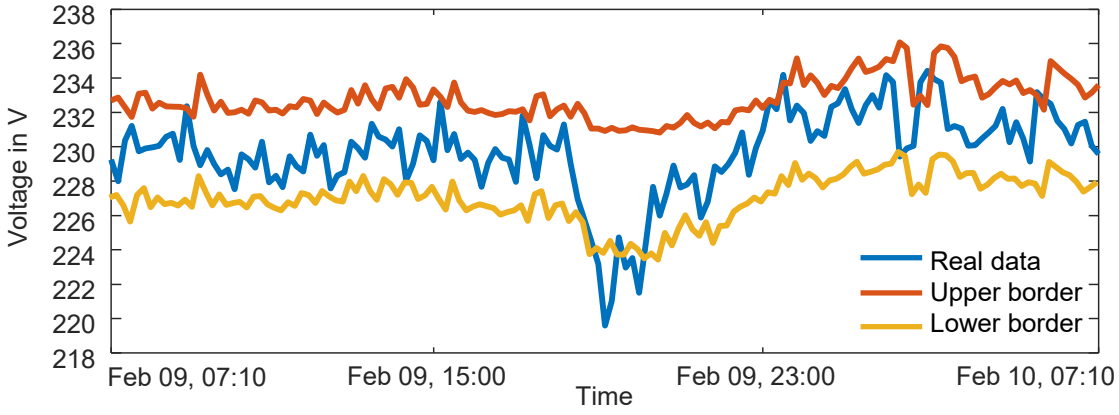


b) the highest error in the upper voltage border (PWCF+NN method)

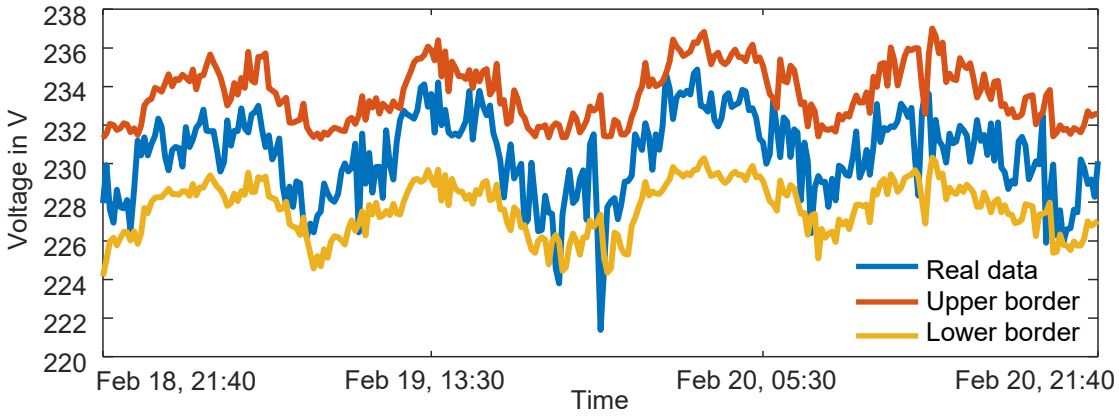
Figure 5.5: Results of the PWCF + NN method applied to voltage data for upper borders. (a) highest yearly voltage across nodes, (b) maximum error in upper border forecasting

In comparison to the simple PWCF method, the boundaries generated by the PWCF+NN approach exhibit reduced smoothness. This behavior is attributed to the neural network's ability to incorporate data fluctuations into the learning process, thereby improving its responsiveness to dynamic system changes. While this adaptation enhances the method's ability to react to real-time variations, it results in less smooth forecast boundaries than those produced by the PWCF-based forecasts.

Note: The results include data from all nodes, which means some simulation outcomes may occur on the same date but show varying values and behaviors. When identifying the highest or lowest values, the specific days may differ, as such extremes can occur multiple times on different dates. To evaluate each method, the worst-case scenario is selected by focusing on instances where forecasting errors associated with extreme values are significant.



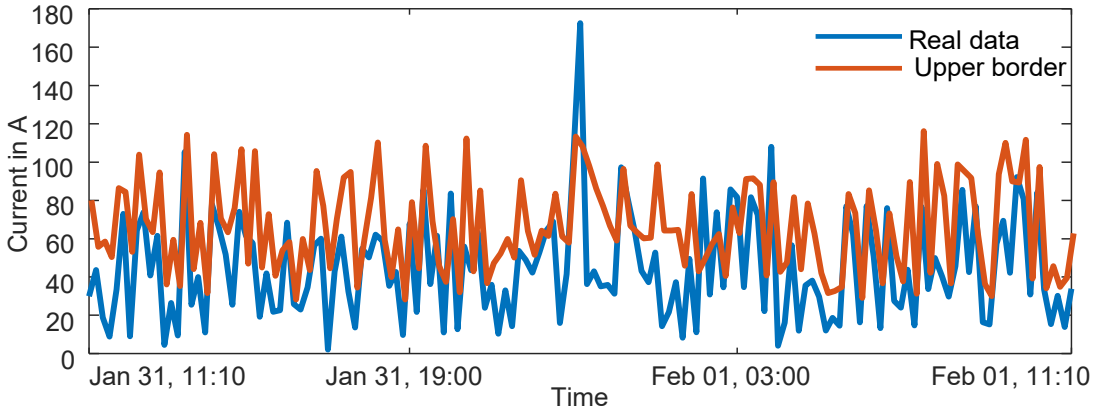
a) the lowest voltage value throughout the year (PWCF+NN method)



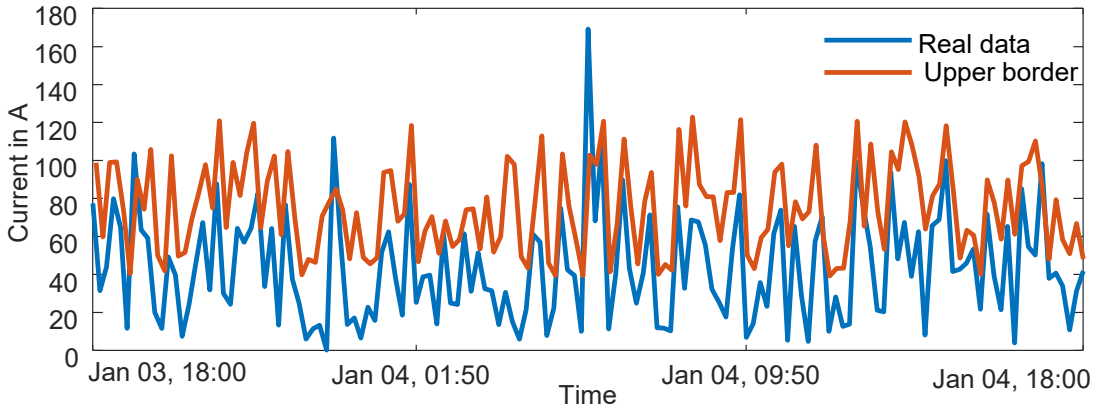
b) the highest error in the lower voltage border (PWCF+NN method)

Figure 5.6: Results of the PWCF + NN method applied to voltage data for lower borders. (a) lowest yearly voltage across nodes, (b) maximum error in lower border forecasting

The PWCF + NN method aims to uncover patterns between inputs and outputs. However, the high fluctuations in real node data make it tough to identify a reliable pattern. As a result, when applied independently, particularly in real LV grids, the NN method tends to demonstrate limited accuracy. This limitation is especially evident in short-term forecasting at the level of individual nodes, which in this study have a sampling interval of 10 minutes. Forecasting at such high temporal resolution for each household is often both complex and impractical. In contrast, focusing on worst-case scenarios has proven to be a more effective and application-oriented strategy. Given the unique forecasting needs in LV grids and the limitations posed by real data, the PWCF method has proven to be a promising solution. However, in its simple form, the method lacks the flexibility to adapt to new or evolving grid environments. By integrating the PWCF method with an NN, a more intelligent and adaptive system has been developed, one that retains the strengths of the original approach while significantly enhancing its adaptability to dynamic grid conditions.



a) the highest current value throughout the year (PWCF+NN method)



b) the highest error in the upper current border (PWCF+NN method)

Figure 5.7: Results of the PWCF method applied to current data (a) highest yearly current across nodes, (b) maximum error in upper border forecasting

The results show that this combined method (PWCF + NN) achieves improved success rates for forecasting both the upper and lower voltage boundaries in the SMAPO dataset. In this context, the success rate (%) is defined as the proportion of forecasts correctly covering the real measured values (Success rate (%)=(1-$\frac{\text{Number of error}}{\text{Total samples}}$) ×100). The forecasting success rate represents the percentage of data points, available total samples, for which the observed values are successfully enclosed by the forecasted upper and lower bounds.

A comparison between the PWCF and PWCF + NN methods is presented in Table 5.3. While the success rate has slightly improved, the primary advantage of the PWCF + NN method lies in its enhanced adaptability to changing grid conditions, particularly in future scenarios. Furthermore, this method exhibits a reduced dependency on the operator's expertise and prior knowledge of the system. It is important to note that methods rely solely on historical data. This reliance limits their ability to accurately forecast extreme values, which are often driven by sudden changes in the system. Such changes are typically triggered by control actions in both LV and medium-voltage grids, such as transformer tap changer operations.

In future work, the integration of data related to these control commands could significantly improve forecasting accuracy. Incorporating this additional information would enable the models to more effectively capture dynamic system behavior, thereby supporting more robust and adaptive grid management strategies.

Table 5.3: Comparison of PWCF and PWCF + NN methods

Features	PWCF	PWCF + NN
Upper voltage border forecasting success rate	98.38	98.73
Lower voltage border forecasting success rate	98.22	98.42
Upper current border forecasting success rate	98.17	98.19
Ability to adapt to new grid environments	Limited flexibility	Enhanced flexibility through NN integration
Response to sudden changes (e.g., transformer tap changers)	Limited responsiveness	Improved responsiveness due to NN integration
Dependency on the operator's knowledge	High (requires operator interpretation of historical data)	Reduced (NN minimizes reliance on operator interpretation)
Data used	Relies on historical data	Combines historical data with NN modeling

In these simulations, the PWCF boundaries are forecasted every 10 minutes, corresponding to the smart meter sampling rate in the SMAPO dataset. At each time step, the forecast is generated for the subsequent interval, thereby providing a continuously updated forecasting of the voltage boundaries. These results can support use cases such as preventive control, where forecasts are utilized to dynamically adjust the sensitivity of control thresholds, enabling earlier and more targeted interventions.

Further analyses were conducted on the SMAPO grid to evaluate the impact of PV systems, the number of considered historical data, and sampling rates on these methods.

- Approximately 12% of households in the grid are equipped with PV systems with an average installed capacity of 5.75 kW, unevenly distributed across phases, with nearly 50% on L1 due to single-phase connections of smaller systems. A comparison between PV and non-PV nodes revealed identical success rates in predicting results. These

findings suggest that PV systems do not represent critical points in this grid. This conclusion may differ in grids with higher PV penetration, particularly when local PV generation exceeds load demand.

- The success rate increases with the number of historical days considered. In this study, a value of D = 28 was selected, as it offers high accuracy while further increases in D yield only marginal improvements and lead to significantly higher computational costs.
- The data sampling rate has a considerable impact on forecasting performance. The methods were tested at intervals of 10, 20, 30, and 60 minutes. Results show a consistent decline in success rates with increasing sampling intervals. The reduction in forecasting accuracy was consistent across all tested thresholds, indicating a uniform sensitivity to sampling rate. However, it is important to note that at higher sampling intervals, the most recent critical information becomes outdated more quickly, further affecting the model's performance.

5.2.2 Results of Anomaly Detection with Artificial Intelligence Methods

ML methodologies were evaluated using real data (AISOP dataset). For the analysis of the methods presented in the sections 4.3.1 to 4.3.5, one urban station, referred to as Station 1, and described in section 5.1.2, was selected as the test case. Given that the current LV grid presents relatively few operational challenges, the primary focus was placed on anticipated future scenarios that may arise due to increased integration of PV systems and additional electrical loads. To accurately assess the performance of the ML-based anomaly detection methods, tailored future scenarios were introduced into the actual 2022 net power data. This approach enabled the simulation of plausible future conditions within the German LV grids. The scenarios were carefully selected to depict minimal but fully observable and detectable changes in net power patterns.

The resulting dataset, referred to as "2022 + added scenario", was constructed by combining real measurement data from 2022 with synthetic modifications representing projected grid developments. For historical context, data from 2017 to 2021 were also included. In the scenario, the addition of four PV systems, each with a capacity of 15 kW_p, three electric vehicles, and four electric heat pumps was modeled. These additions were specifically made during two periods in 2022: from January 1 to February 12, and from July 28 to September 7. An overview of these additions is provided in Figure 5.8.

The Seasonal-Trend decomposition using the LOESS[11] (STL) method is applied to decompose the time series into trend, seasonal, and residual components. By separating recurring seasonal patterns from the underlying trend, STL enables a clearer interpretation of data behavior and facilitates the identification of both long-term developments and short-term

[11] locally weighted regression and scatterplot smoothing

anomalies. This decomposition is particularly useful for DSOs, as it supports both operational control and long-term planning by highlighting significant patterns and filtering out minor fluctuations. By isolating and removing seasonal effects, such as daily and weekly cycles, and extracting the underlying trend, STL provides DSOs with a clearer and more actionable interpretation of the data.

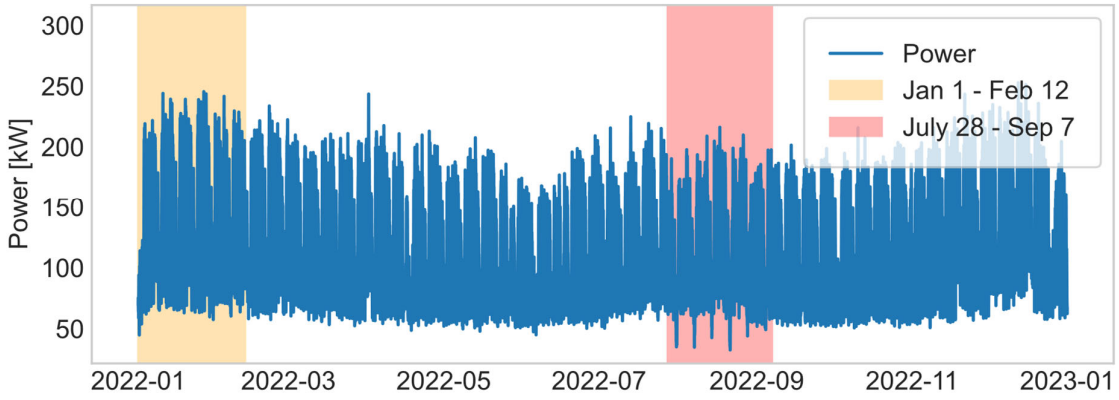


Figure 5.8: Two added scenarios in real data (2022 + added scenario)

Figure 5.9 illustrates the trend, the daily (seasonal 1) and weekly (seasonal 2) seasonal components, and the residuals obtained from the STL decomposition of the net power data at the selected station. The seasonal components are centered around zero, ensuring that they capture only deviations from the underlying trend. Consequently, seasonal values can be both positive and negative, where negative values indicate that the net power lies below the trend at the corresponding time step.

Figure 5.10 presents the results of applying the STL method for anomaly detection to the trend and residual components of the net power time series. Figure 5.10 (a) highlights significant trend anomalies, such as meter failures in October 2019, pronounced changes in behavior pattern in April 2020 due to the coronavirus pandemic, and shifts observed in August 2022 due to the extra PV scenario. In contrast, Figure 5.10 (b) focuses on anomalies in residuals, which reflect minor variations in the data. For example, Figure 5.11 (a) details anomalies identified in both the trend and residual components throughout 2018, with specific causes attributed to each. Residual anomalies observed on January 1st are associated with atypical energy consumption patterns characteristic of New Year's Day, while trend anomalies detected in early March correspond to a notable increase in power values.

A further investigation of these variations is presented in Figure 5.11 (b), which relates the identified anomalies to global solar radiation and ambient temperature over the corresponding period. During this week, ambient temperatures dropped to approximately −10 °C, leading to increased electricity demand, while periods of reduced solar irradiance limited PV generation. The combined effect of elevated load demand and reduced local generation is consistent with the observed anomalies in net power consumption.

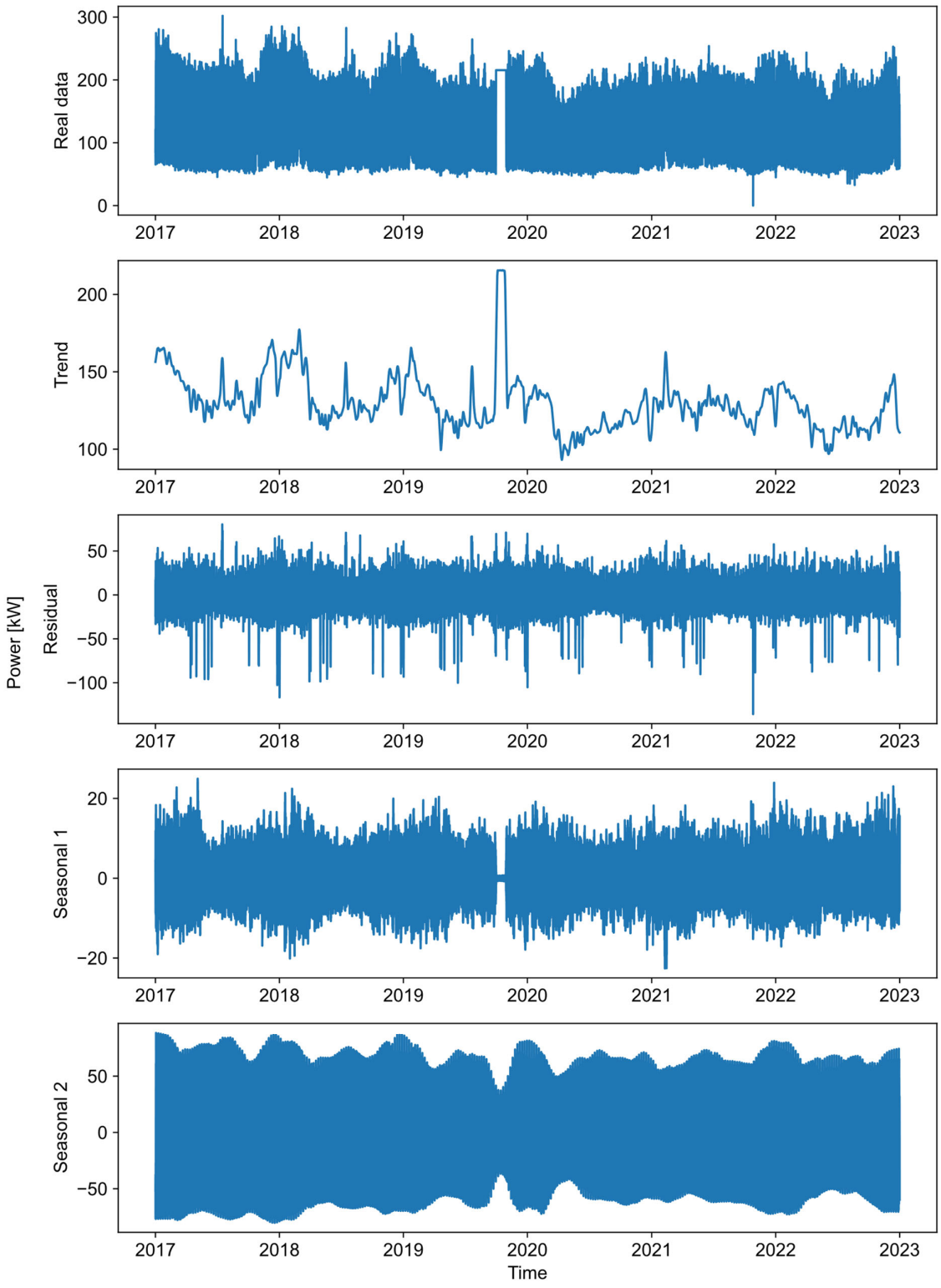
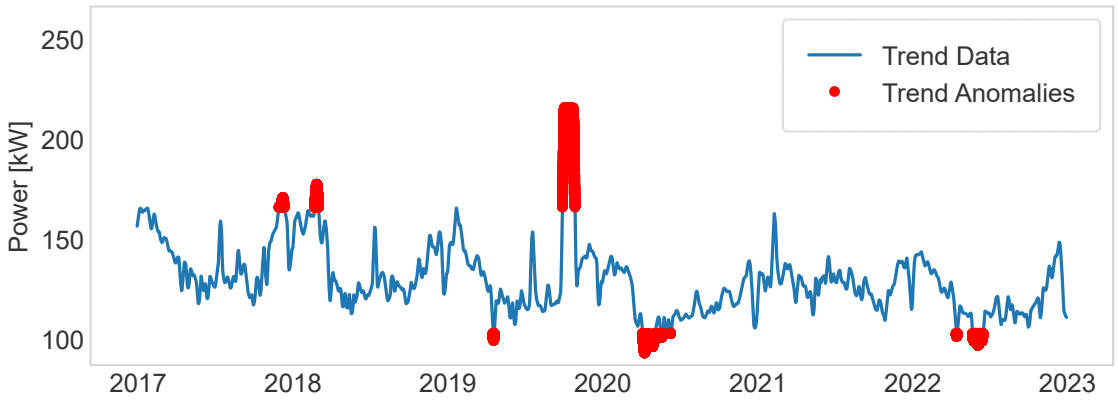
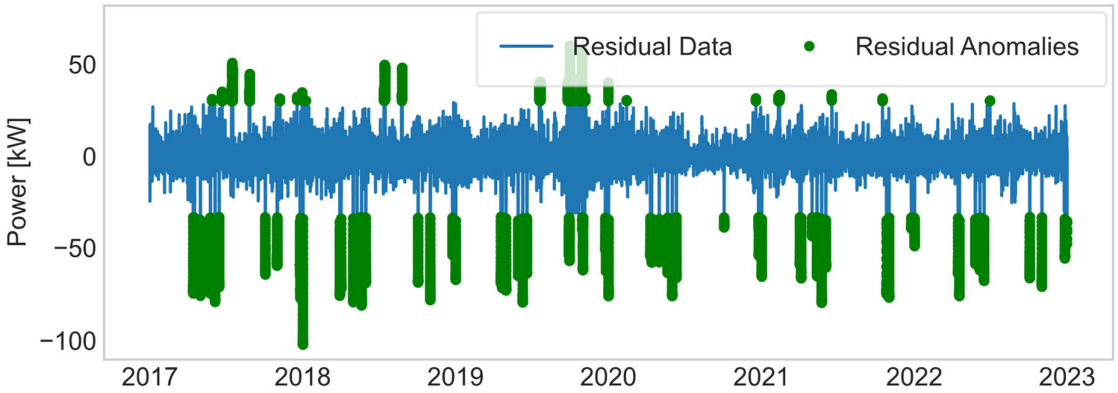


Figure 5.9: STL components for the net power of station 1



a) Trend anomalies in net power detected by the STL method



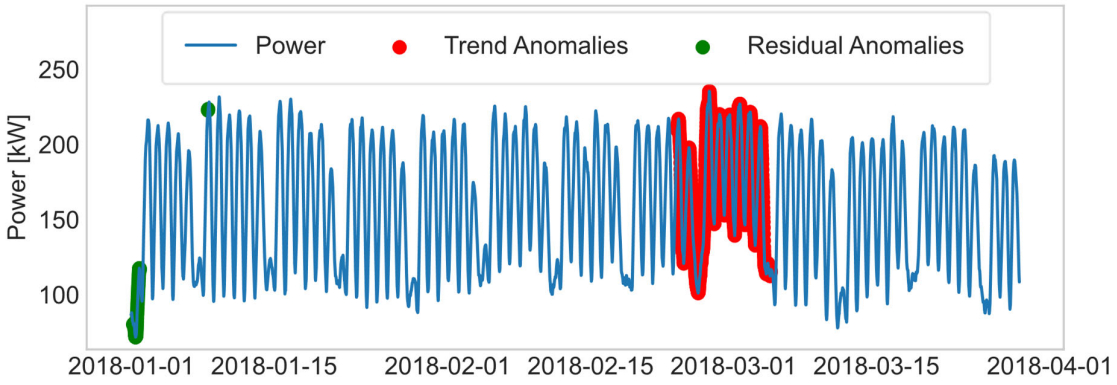
b) Residual anomalies in net power detected by the STL method

Figure 5.10: Anomaly detection in net power using the STL method

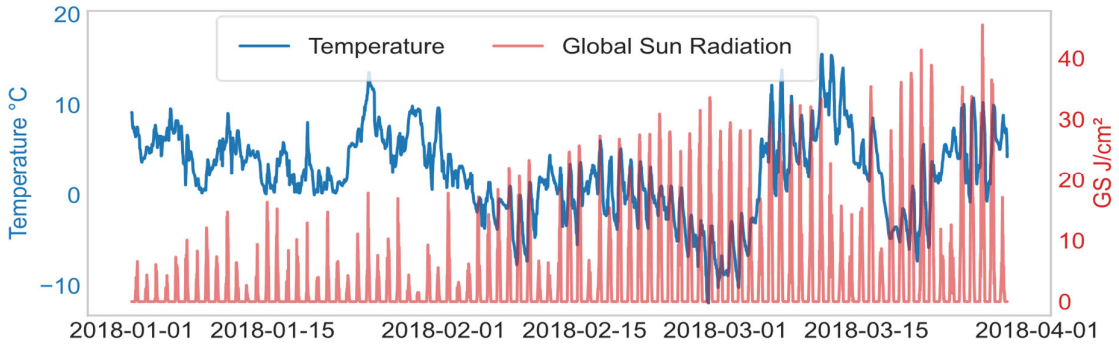
The utility of the STL method also extends to its application on live data, as demonstrated in Figure 5.12, the recent data analysis from April 1 to October 10, 2024. which presents the analysis of measurements collected between April 1 and October 10, 2024. This application illustrates the method's effectiveness in real-time environments, underscoring its robustness and practical relevance for ongoing utility monitoring and management, including voltage measurements across all three phases. Compared to a simple threshold-based check (e.g., U> limit), the STL method provides a significant added value by decomposing the signal into trend, seasonal, and residual components. This allows it to distinguish between natural cyclic variations and true anomalies, reducing false alarms caused by regular load patterns or measurement noise. This demonstrates the method's versatility and effectiveness in analyzing complex electrical datasets.

This method provides a clear and interpretable decomposition of complex power data, enabling operators within DSOs to gain a comprehensive overview of system behavior. By isolating the long-term trend and removing short-term fluctuations and seasonal noise, STL decomposition facilitates more reliable monitoring and interpretation of power flow dynamics. This is particularly beneficial for anomaly detection, as it allows operators to focus on meaningful deviations from expected behavior rather than being distracted by routine variability.

Consequently, the STL method supports informed decision-making in operational planning and the integration of renewable energy sources within low-voltage grids.



a) Trend and residual anomalies in 2018



b) Correlation of net power anomalies with temperature and GS

Figure 5.11: Detailed analysis of net power anomalies correlated with environmental factors

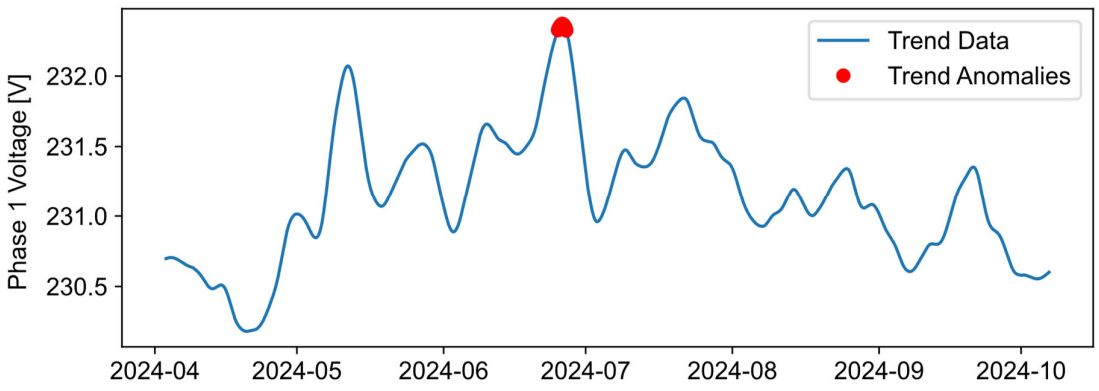


Figure 5.12: Trend anomaly detection by the STL method for voltage

For the assessment of future scenario impacts at a rural station, the Long Short-Term Memory (LSTM) Autoencoder method was employed to analyze data from station 2. In Figure 5.13, the actual net power consumption for 2021 is presented alongside projections that incorporate

scenarios with the addition of ten electric heat pumps (EHPs) and four additional PV generation units throughout the entire year. These projections highlight how power dynamics might be altered due to increased energy demands and generation capacities.

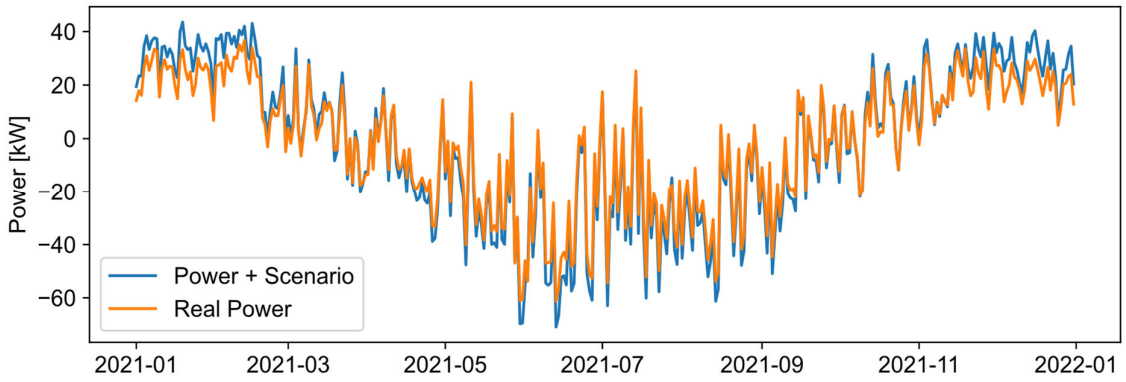


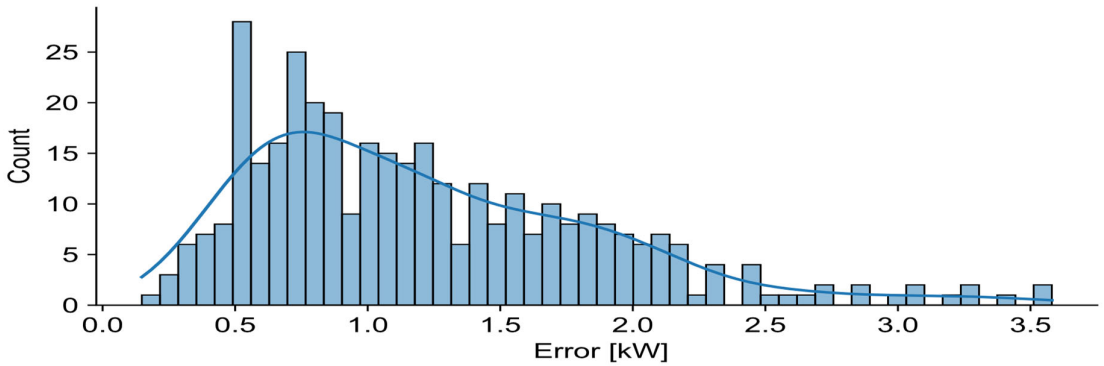
Figure 5.13: Comparison of net power in “real net power 2021” and “real net power 2021+ added scenario

Data from 2017 to 2020 were used as historical training data, while data from 2021, including the added scenario, were reserved for testing and validation. The LSTM models were trained using backpropagation with learning rates between 0.001 and 0.01, optimized via the Adam algorithm ($\beta_1 = 0.9, \beta_2 = 0.99, \in= 10^{-7}$) with a batch size of 64 and 100 training epochs.

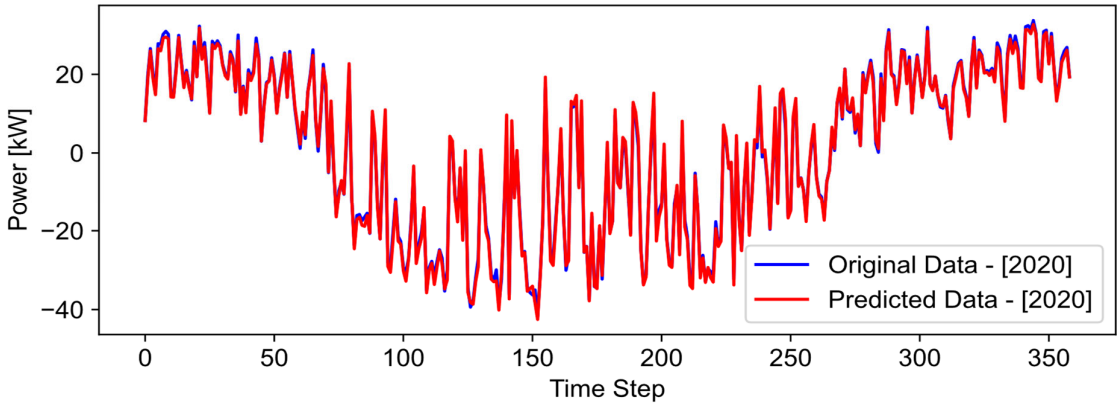
Figure 5.14 presents the reconstruction errors and the model output when applied to the 2020 training data. The median reconstruction error is 1.08 kW, while the 95th percentile and maximum reach 2.46 kW and 3.58 kW, respectively. This indicates that the model accurately reproduces the majority of training sequences, with only a few outliers showing higher deviations.. However, as shown in Figure 5.14 (b), the largest errors occur during periods when the potential of PV is exceptionally high. This observation highlights a specific challenge encountered in rural stations, where increased PV generation can introduce greater variability and forecasting difficulty.

The data for 2021 are considered test data for evaluating the method, including the added future scenario, which is presented in Figure 5.13. The 2021 dataset presents a particular challenge, as its trend deviates significantly from previous years, suggesting a notable increase in PV installations. The added scenario highlights this trend, and while the predicted values for 2021 show reduced accuracy compared to earlier years, the model effectively tracks the trend for most non-summer months, as shown in Figure 5.15(a).

Figure 5.15 (b) illustrates the anomalies detected within the 2021 data, where high anomaly scores, indicating significant deviations from expected normal operational patterns, are marked in red, and lower scores, indicating minor deviations, are marked in yellow. The anomaly score represents the magnitude of deviation between the actual and predicted power consumption values.



a) Reconstruction errors for training data



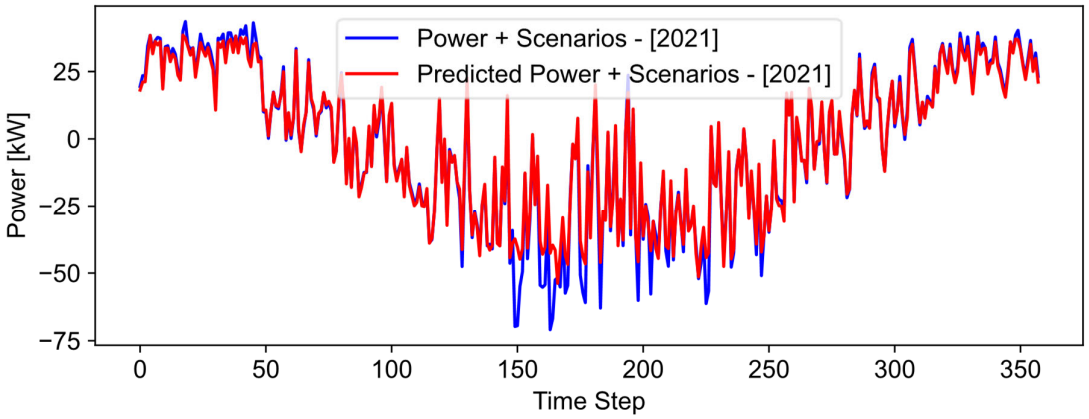
b) Training output of net power data for 2020

Figure 5.14: Analysis of Net Power Data for 2020

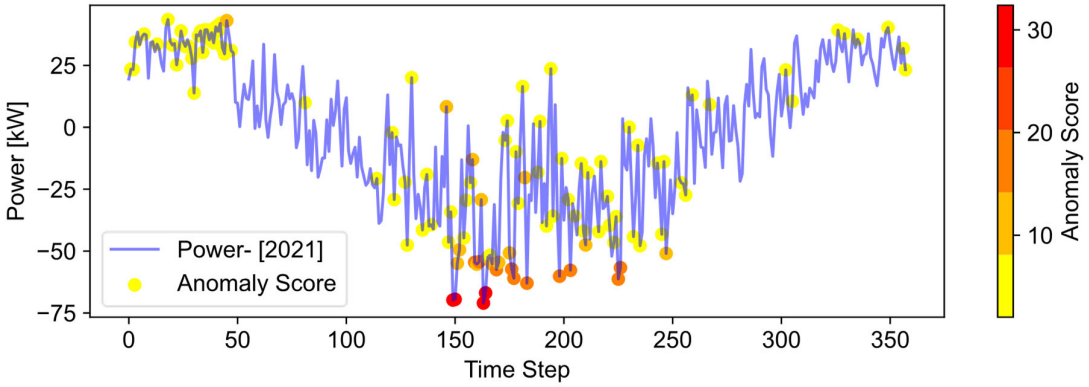
These results demonstrate the method's capability to identify additional PV installations and electric heat pumps in the dataset. Notably, the error clusters reveal that the model's predictions exhibit higher losses during the warmer months (March to September) compared to the colder months (October to February). This discrepancy is attributed to the increased aggregate PV power output during summer, which introduces greater variability and poses additional challenges for accurate forecasting.

Figure 5.16 presents the clustering of reconstruction errors for 2021 using the K-Means algorithm, where the errors are grouped into three distinct clusters: low (cluster 0), medium (cluster 1), and high (cluster 2). This clustering approach proves particularly valuable for identifying patterns in the model's performance and isolating periods of high variance. By categorizing errors into meaningful groups, it becomes easier to identify when and under what conditions the model's predictions deviate significantly from the actual values.

The advantage of clustering lies in its ability to highlight specific periods or conditions that cause substantial deviations. High-error clusters, for instance, may indicate potential issues such as unaccounted PV installations, anomalies introduced by the added scenario, or seasonal effects that increase variability, particularly during summer months when PV generation is at its peak.



a) Power values for 2021 with scenarios and predicted results



b) Anomalies for power values in 2021 with future scenarios

Figure 5.15: Analysis of net power data for 2021 test scenario

This targeted analysis enables a more systematic investigation of model limitations, facilitates adjustments to improve prediction accuracy, and helps identify real-world changes in the data that the model may have struggled to capture. Based on the analysis of anomalies, clustering results, and additional information such as global solar radiation, weather conditions, and seasonal patterns, the reason for anomalies in each cluster can be distinguished. While this method does not provide absolute certainty, it offers valuable insights into the probable drivers of observed anomalies. As shown in Figure 5.17, cluster 0 contains anomalies that mainly occur under medium-low global solar radiation conditions (53%), medium-low temperatures (48%), and mainly during the cold months (October to February, 56%). This pattern suggests that many anomalies in this cluster are likely driven by increased power consumption due to extra energy demand from electric heat pumps.

In contrast, cluster 2 shows a high concentration of anomalies during periods characterized by medium-high to high solar radiation and medium-high to high temperatures. These anomalies primarily occur during the warmer months, when increased PV generation introduces greater variability and challenges for accurate forecasting. This pattern indicates that anomalies in Cluster 2 are likely related to the high output potential of PV systems under sunny conditions,

conditions that may not be fully captured by the model or the parameters defined in the added scenario.

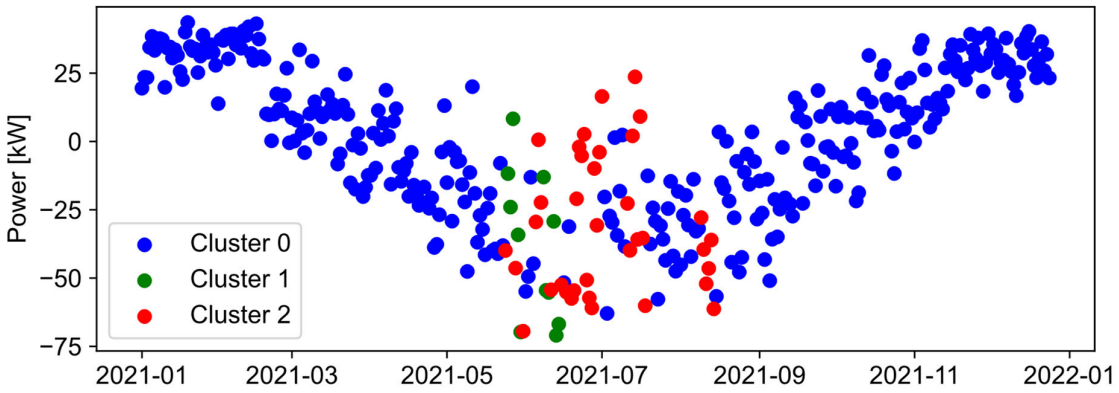


Figure 5.16 Clustering of reconstruction errors in anomaly detection

As shown in Table 5.4, each anomaly detection method can be used for different use cases in grid management and control. For instance, trend-based anomaly detection, such as the Isolation Forest applied to trend components, is particularly effective for identifying deviations influenced by trends, providing critical insights for grid monitoring. By detecting anomalies related to changes in net power, grid operators can analyze grid sensitivity to environmental conditions, operational demands, and other dynamic factors.

Through feature sensitivity analysis, operators can determine how variables like temperature, solar radiation, and consumer behavior impact grid performance. For example, high sensitivity to weather conditions highlights the need for adaptive operational strategies in regions reliant on solar and wind energy. Additionally, consumer engagement can be enhanced by informing users how their consumption patterns affect operational performance, particularly during peak demand periods, and encouraging more efficient energy use.

This highlights the importance of choosing the appropriate anomaly detection method based on specific analytical objectives, whether it’s detecting local anomalies, monitoring trends, or understanding grid sensitivity, ensuring robust and efficient grid control.

The results of the other anomaly detection methods introduced in Chapter 4 (sections 4.3.1 - 4.3.4) are presented in Appendix F.

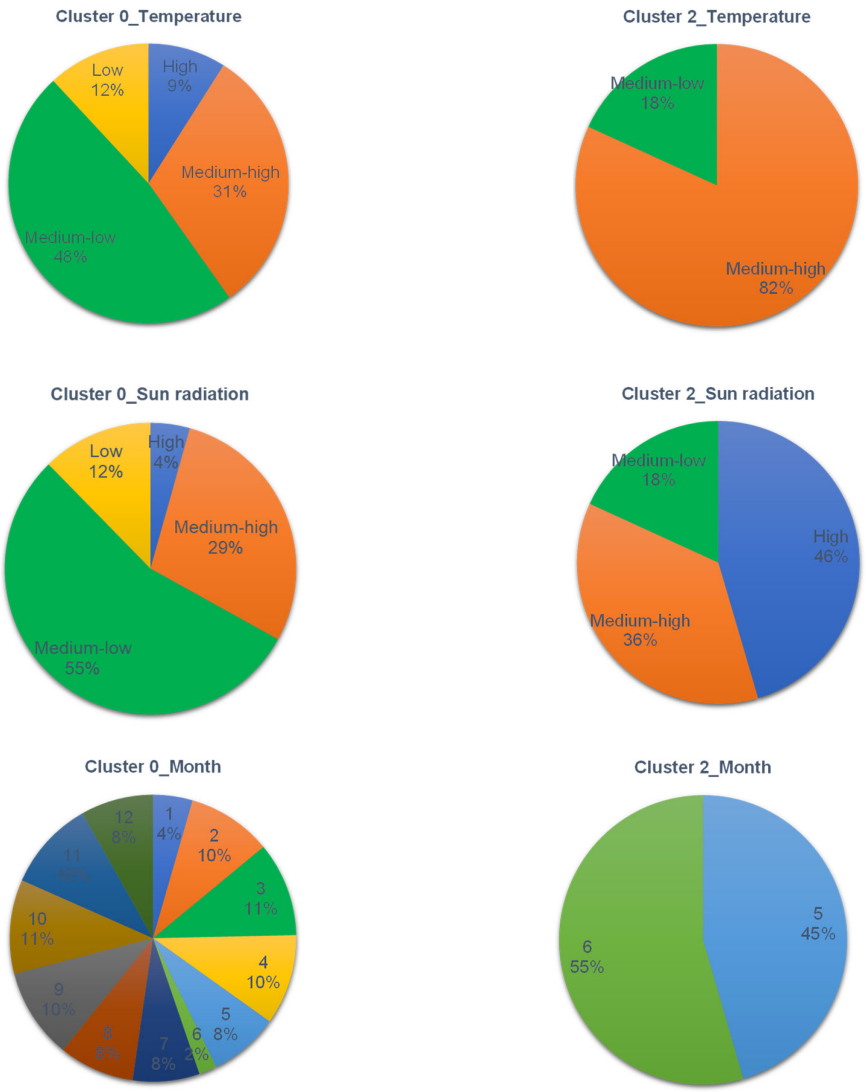


Figure 5.17: Analysis of anomaly characteristics for two clusters

Table 5.4: Comparison of AI methods for anomaly detection.

Approach / Characterization	LSTM Autoencoder	Isolation Forest for Trend and Residual	Correlative Isolation Forest	Statistical Windowing with Isolation Forest	Extreme Studentized Deviate	Cluster-Enhanced Feature Regression
Sensitivity to Small Changes	High	Trend: Moderate Residual: High	Moderate	Moderate	High	Moderate
Dependency on Parameters	Moderate	Low	Low	Low	High	Moderate
Anomaly Detection Based	Reconstruction errors in latent space	Decomposed trend and residual components	Correlation with historical windows	Local patterns in data windows	Extreme deviations	Feature deviations from expected patterns
Typical Data Used	Power values	Power values	Power values, temperature, solar radiation, season, wind	Power values, temperature, solar radiation	Power values	Power values, temperature, solar radiation, weekdays
Typical Purpose	Monitoring and planning	Monitoring and planning	Analysis	Monitoring and analysis	Analysis	Monitoring and analysis

5.2.3 Results of Anomaly Detection with Statistical Method

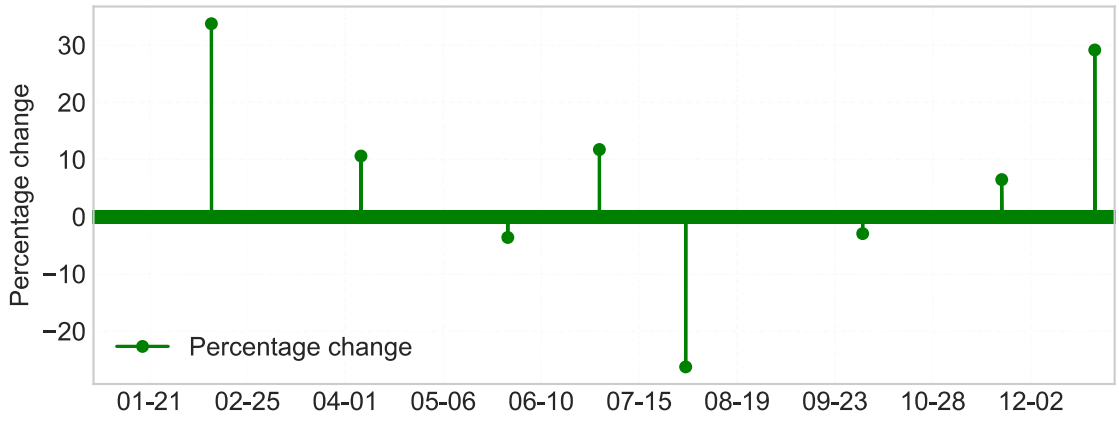
Statistical methods offer the advantage of requiring limited model training while providing interpretable results based on historical behavior patterns and distributional shifts. In the following, the results of the three statistical anomaly detection approaches presented in Section 4.3 are applied to the dataset from Station 1, which represents an urban station.

The probabilistic normalized expectation ratio cumulative sum (CUSUM) method (introduced in Section 4.4.1) is applied to identify abrupt changes in the net power time series by monitoring deviations from expected behavior over time. Figure 5.18 (a) shows the percentage change relative to the Frequent Behavior Weighted Mean (FBWM) pattern, indicated by the green shift marker. In Figure 5.18 (b), the blue line represents net power values in 2022, the dotted green line indicates change points, and the red shading highlights the probability of changes. These probabilities quantify the likelihood that a significant shift in the mean value of the signal has occurred. When the probability exceeds the predefined limit (p_{limit} = 0.001), the event is classified as a change point, indicated by a green dashed line.

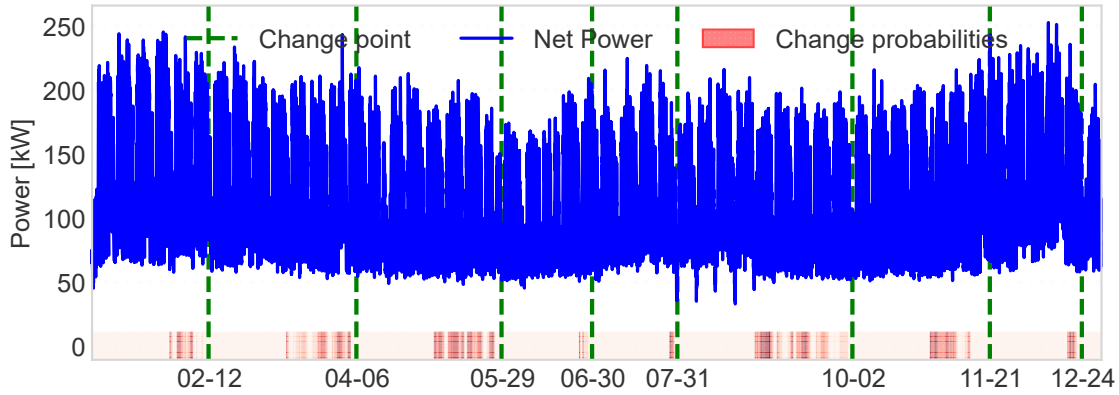
A series of change points on February 12 corresponds to the end of the added scenario, reflecting increased energy consumption due to the integration of additional electric vehicles and electric heat pumps. Additional change points observed at the end of July align with the increase in PV power generation resulting from the installation of four additional PV units. These results demonstrate the statistical method's capability to detect significant shifts in power consumption and generation, thereby validating the impact of the modeled scenarios on the net power profile.

The following results present the seasonality analysis performed using the probability density function (PDF) method (introduced in 4.3.2), designed to detect weekly seasonality patterns in power data and identify changes or anomalies in these patterns over time. This is achieved by first isolating the weekly seasonality using the STL (Seasonal-Trend decomposition using LOESS) method, which separates the time series into trend, seasonal, and residual components. By isolating the weekly seasonality using the STL method, recurring variations specific to weekly cycles are extracted, independent of long-term trends and short-term noise.

Figure 5.19 illustrates how the weekly seasonality patterns have evolved over different years due to external factors. Specifically, Figure 5.19 (a) shows the pattern of 2021 during Week 4, which exhibits a distinct shape corresponding to the pandemic period, reflecting the unusual conditions and changes in load behavior during that time. Figure 5.19 (b) highlights the pattern of 2020, which was similar to other years in Week 4 but changed significantly by Week 13, the first week of the pandemic, indicating the impact of the sudden shift in consumption patterns. Lastly, Figure 5.19 (c) shows the pattern of 2022, where the increased penetration of PV generation results in a significantly altered weekly profile. This change highlights the influence of distributed energy resources on the structure of weekly seasonality.



a) Percentage change over time



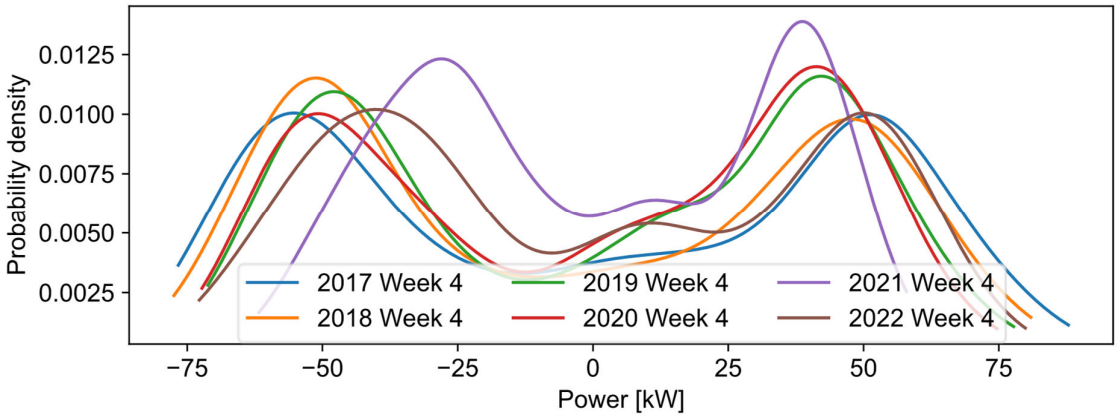
b) Detected change points and corresponding change probabilities

Figure 5.18: CUSUM results for net power in 2022 to detect change points

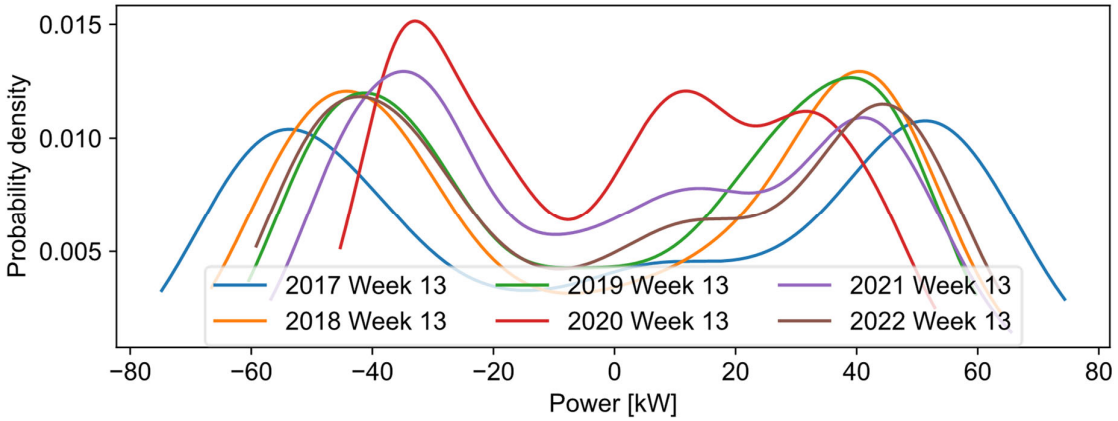
The weekly resolution was selected because both weekday, weekend pattern, and short-term generation patterns (e.g., PV output fluctuations) can be captured and compared across years. From an operator's perspective, weekly seasonality analysis provides actionable insights for short-term operational planning. The ability to compare these PDFs across years highlights the method's effectiveness in detecting changes in weekly patterns. Shifts in the peaks, changes in the range, or alterations in the shape of the PDFs indicate potential deviations in the underlying seasonal behavior.

Figure 5.20 compares weekly seasonality patterns of consecutive years using the Kullback-Leibler Divergence (KLD) method, where higher KLD values indicate greater differences. Key anomalies are evident: between 2018 and 2019, a significant spike in KLD during Weeks 39 to 44 corresponds to a meter failure from September 25, 2019, to October 30, 2019, causing abnormal pattern changes. In 2020, compared to 2019, an increase in KLD after Week 13 reflects the impact of the COVID-19 pandemic, which disrupted consumption patterns. Comparisons for 2021 and 2022 show smaller deviations, with patterns stabilizing post-pandemic, though changes due to increased PV generation in 2022 are apparent. This

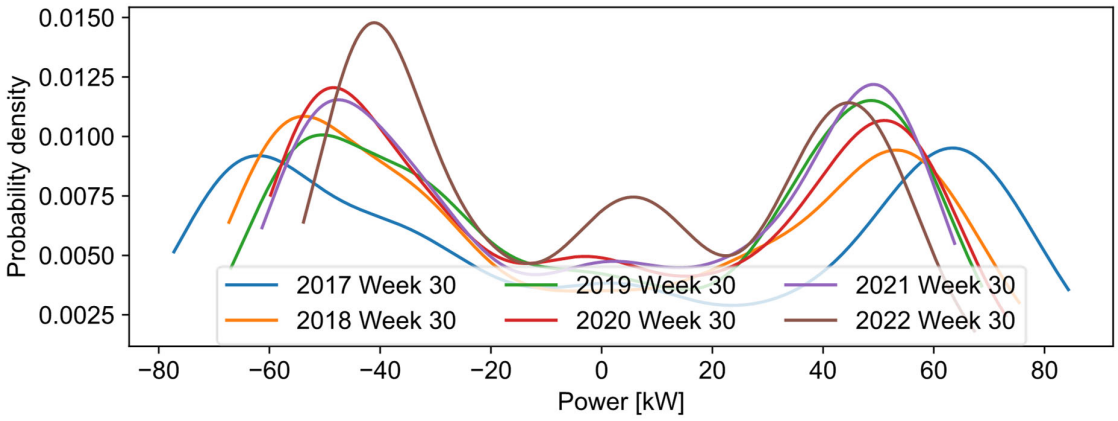
analysis highlights KLD's ability to detect anomalies and external impacts on grid behavior.



a) PDFs of weekly seasonality patterns for net power in week 4



b) PDFs of weekly seasonality patterns for net power in week 13



c) PDFs of weekly seasonality patterns for net power in week 30

Figure 5.19: The PDFs of the weekly seasonality patterns across several years

By relying solely on busbar data, operators can monitor recurring patterns and detect significant changes without requiring extensive sensors and meters in the grid. Operators can use this technique to identify gradual changes in weekly load patterns, allowing proactive

adjustments in grid operation strategies. It is also useful for evaluating the impact of external factors, such as increased integration of PVs, and for identifying seasonal grid vulnerabilities.

This approach not only enhances the ability of operators to monitor and maintain LV grids but also supports the proactive management of grid stability, particularly as grid dynamics evolve with increasing distributed energy resources integration and changing load behaviors.

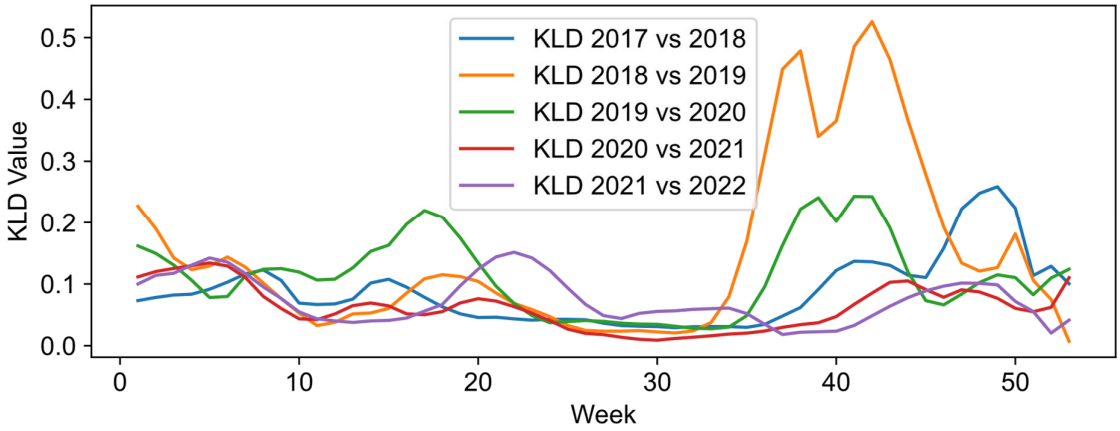


Figure 5.20: KLD-based comparison of weekly seasonality across years

The dynamic threshold clustering method (introduced in section 4.4.3) detects anomalies by comparing real-time power values with dynamic thresholds that are defined for specific conditions such as temperature, solar radiation, time of day, and day type. During the analyzed period, March 2020, when the first COVID-19 quarantine measures were introduced, most anomalies were detected at the minimum power boundary, as shown in Figure 5.21. The real power data (blue line) alongside the dynamic maximum and minimum power thresholds in March 2020. The lower boundary is crossed at several points, indicating a consistent drop in consumption during this period. The maximum boundary is rarely exceeded, suggesting that the anomalies in this time frame are primarily due to unusually low demand. This reflects a sudden decrease in power consumption, likely due to changes in daily activity patterns.

Figure 5.22 presents the identified minimum anomalies, marked in red, and data points near the minimum boundary (alarm points), marked in yellow. The real power data is plotted to show where it falls relative to these thresholds. A noticeable increase in anomalies can be observed between March 13 and March 25, which aligns with the sharp reduction in power consumption during the initial COVID-19 lockdown phase.

Figure 5.23 displays the distribution of minimum anomalies across different clusters. The most affected clusters include clusters 41, 113, 119, and 103, which correspond to environmental conditions characterized by low to medium-low temperatures, low to medium-high global solar radiation, and morning or afternoon hours. These results indicate that the anomalies are not randomly distributed but are concentrated in specific contextual scenarios.

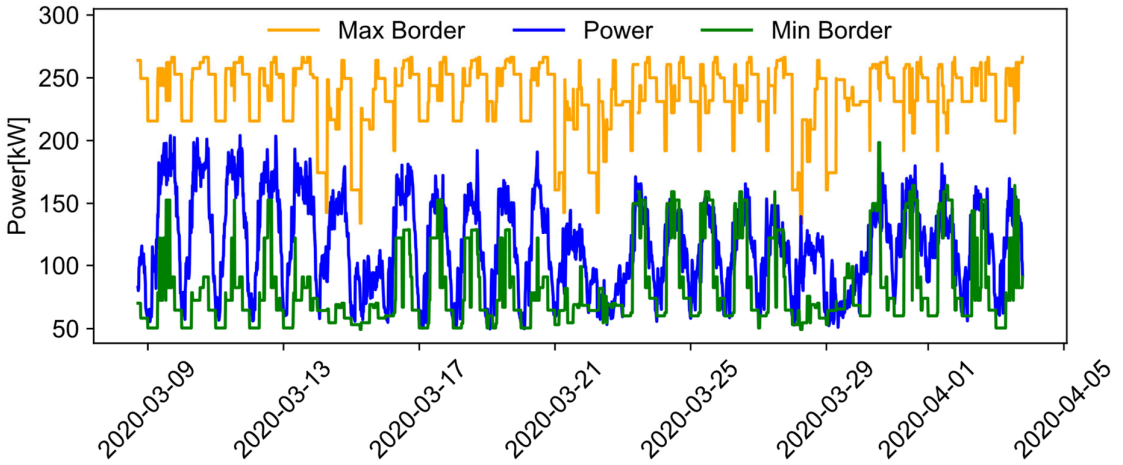


Figure 5.21: Maximum and minimum power thresholds vs. real net power data

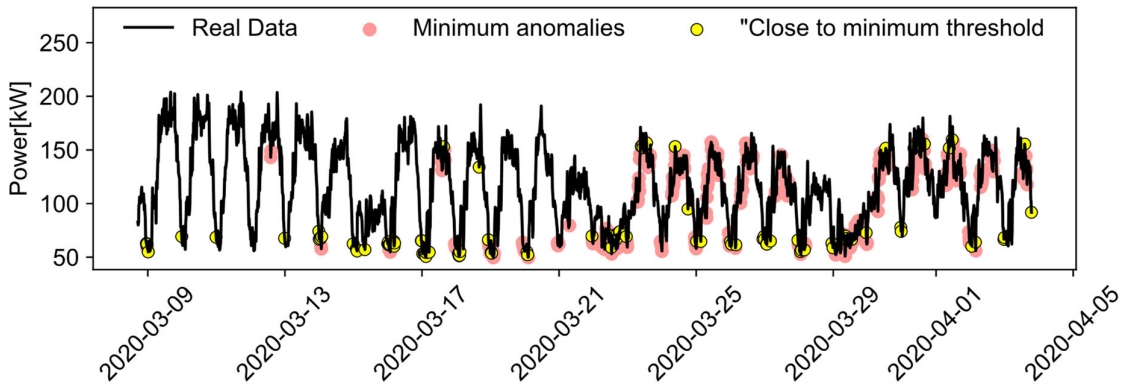


Figure 5.22: Dynamic threshold clustering results for minimum power anomalies during March 2020

Based on the cluster feature Table 5.5, which maps each anomaly to its corresponding environmental condition, it becomes evident that the majority of anomalies occurred under low-temperature conditions. According to the clustering framework introduced in Section 4.4, a total of 128 clusters were generated from the four features: temperature, global solar radiation, hour of the day, and day type. Each cluster represents a unique scenario, allowing precise mapping of anomalies to specific operational contexts.

Table 5.5 analysis and clustering results together show that, in this urban station, most anomalies over the entire dataset occurred during cold temperature periods, regardless of solar radiation levels. This suggests that in urban LV grids, temperature fluctuations have a stronger influence on anomaly occurrence than PV generation. The relatively limited role of PV-related anomalies reflects the lower PV penetration in this particular grid area. Thus, consumption-driven anomalies, particularly during heating-demand periods, dominate the anomaly profile, highlighting the sensitivity of urban load patterns to seasonal temperature

changes.

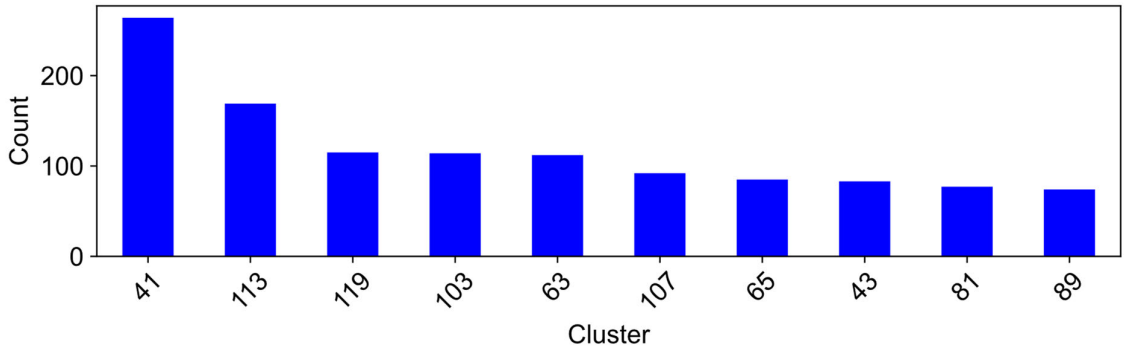


Figure 5.23: Distribution of minimum power anomalies across clusters

Table 5.5: Environmental feature categories for the top three clusters with the highest number of minimum power anomalies

Tempreture category	Global solar radiation category	Hours category	Cluster number
Low	Low	Afternoon	41
Medium-low	Medium-high	Afternoon	113
Medium-low	Medium-high	Morning	119

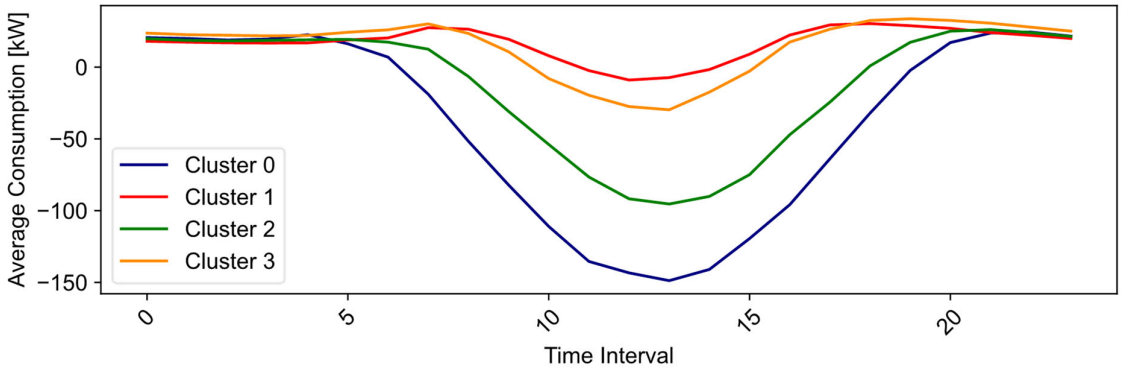
5.2.4 Results of Detecting Unregistered Installed PV

In this method, which is based on clustering analysis, the primary objective is to detect additional installed PV within the grid that represents blind spots in current monitoring systems. For this purpose, the results are presented for Station 2, which represents a rural station with a higher share of installed PV capacity compared to the urban station. The greater PV penetration in this station increases the likelihood of detecting hidden or unmonitored PV systems based on net power behavior. To achieve this, net power data is clustered into four groups based on similar average daily behavior. This clustering approach enables the identification of groups with distinct patterns of net-power changes, which may indicate the influence of PV generation.

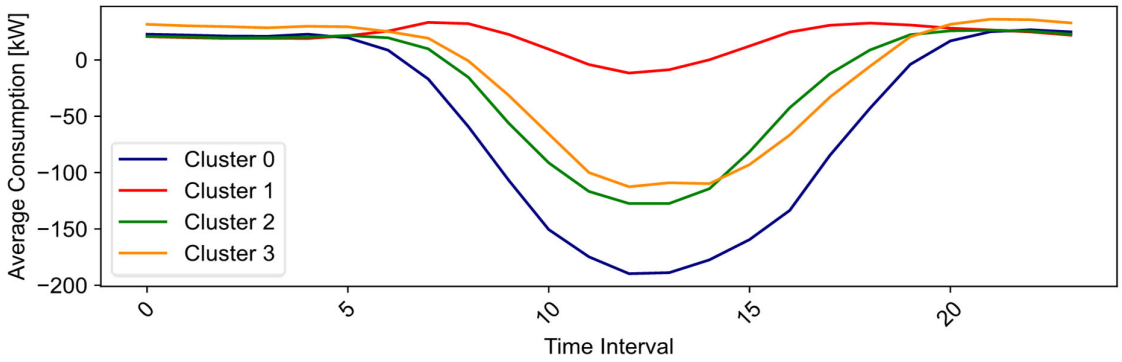
As shown in Figure 5.24, the clustering results for 2021 and 2022 highlight distinct group behaviors. Notably, cluster 0 is characterized by a strong potential for PV generation, where a significant decrease in net power is visible during the day. Figure 5.25 further examines the monthly distribution of these clusters. The majority of members in cluster 0 belong to the warm months (March to September), which corresponds to periods with a high likelihood of PV generation due to increased solar radiation. In contrast, cluster 1 is predominantly associated

with the cold months (October to February), where PV generation is significantly lower, and net power behavior reflects higher energy consumption patterns.

To detect potential extra or unregistered PV installations, the maximum PV generation scenario is applied using realistic assumptions based on panel specifications. In this study, a 72-cell solar panel with a capacity of 250 W is used as the baseline unit. Global solar radiation and time data are used to estimate the maximum possible PV output for each day and location. A graphical user interface (GUI) was developed to support these calculations over multiple years for each station.



a) Clustering of net power data for 2021



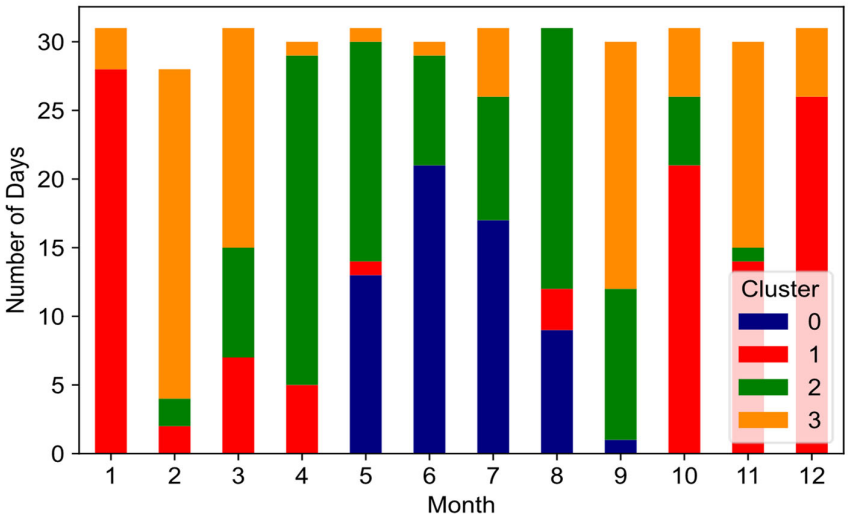
b) Clustering of net power data for 2022

Figure 5.24: Clustering of net power

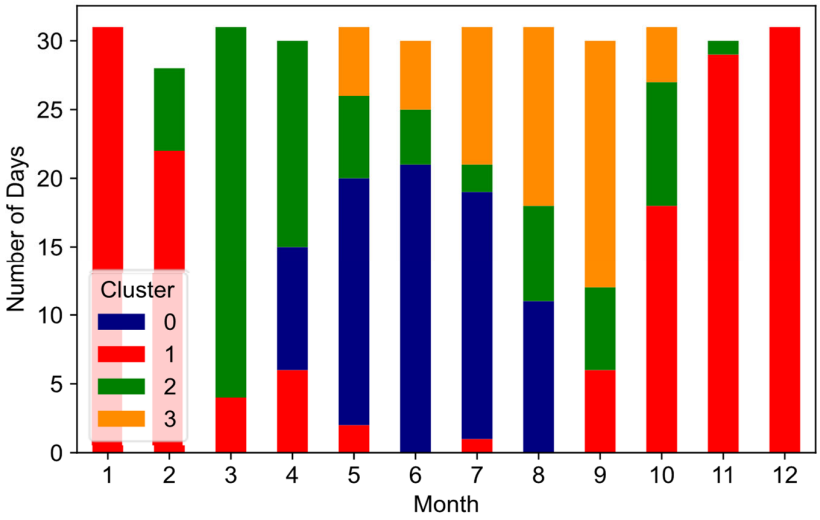
The consumption power is then derived by subtracting the estimated maximum PV generation from the recorded net power. This methodological step effectively isolates the actual load profile by removing the influence of PV generation from the measured data. This distinction is particularly important for anomaly detection, as it ensures that any irregularities observed in the data can be more reliably attributed to issues or variations within the PV system itself.

Figure 5.26 presents the clustering results based on the calculated consumption power for 2021 and 2022. As expected, consumption power values in 2021 remain positive, assuming no battery discharging occurred. However, in 2022, several data points exhibit negative consumption power, suggesting the presence of additional, unmonitored PV installations

feeding into the grid.



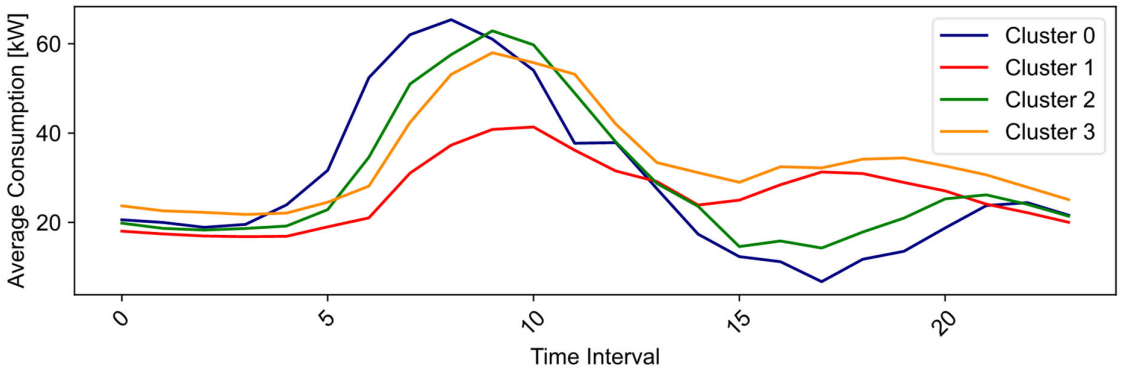
a) Monthly distribution of the clusters for 2021



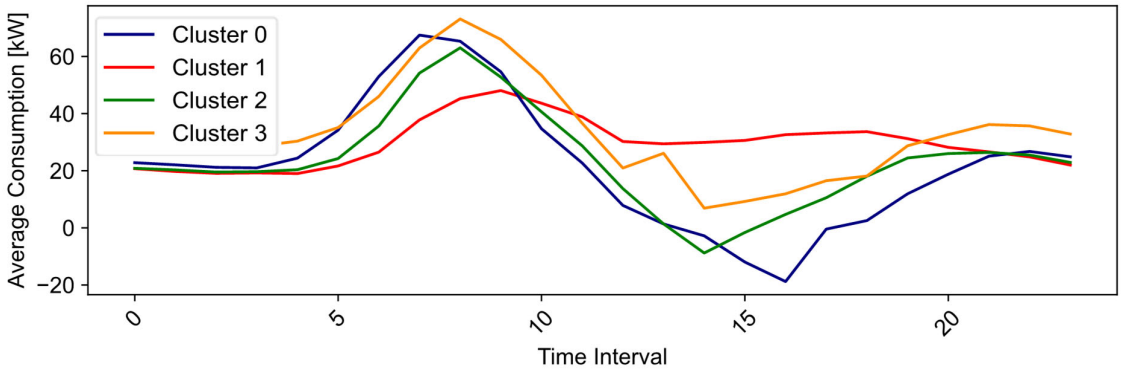
b) Monthly distribution of the clusters for 2022

Figure 5.25: Monthly distribution of the net power clusters

To verify that these negative values are indeed due to PV generation and not influenced by battery discharging or other behaviors, Figure 5.27 presents the monthly distribution and occurrence rate of the identified clusters. The results indicate that the majority of Cluster 0 members, those with strongly negative consumption power, occur during warmer months with higher solar radiation, when PV output is expected to peak. This seasonal correlation confirms that the observed anomalies are most likely caused by additional PV panels rather than alternative sources such as battery storage.



a) Clustering of consumption power data for 2021



b) Clustering of consumption power data for 2022

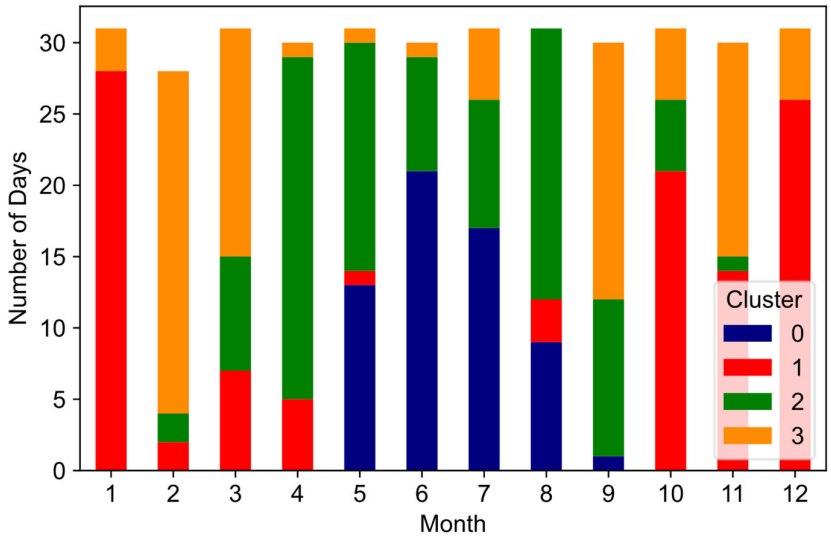
Figure 5.26: Clustering of consumption power

5.2.5 Results of Outage Detection

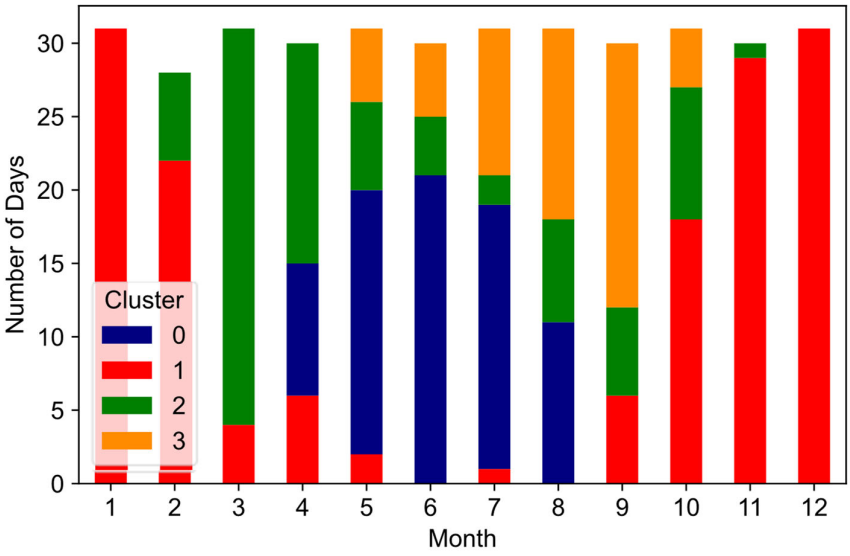
This method focuses on detecting outages in LV grids using measurements taken exclusively at the busbar on the LV side, with no additional data from other nodes within the grid. The method introduced in Chapter 4.6 identifies outages by detecting anomalies in power flow parameters, specifically active and reactive power. To clarify, voltage values are not used for outage detection in this method; the detection mechanism is based solely on deviations in power consumption behavior.

To simulate outages, random outages are generated at different nodes and times within the grid. For each simulated outage, power flow calculations are performed to determine the corresponding active and reactive power values at the busbar. Only these calculated busbar power measurements are considered as input data for the outage detection algorithm, ensuring the method relies solely on busbar-level observations without additional node-specific measurements.

Figure 5.28 illustrates the anomaly detection process based on the differences in power values observed at the busbar under disturbance conditions. The blue line represents the computed differences between consecutive measurements of active power at the busbar.



a) Monthly distribution of the clusters for 2021



b) Monthly distribution of the clusters for 2022

Figure 5.27: Monthly distribution of the consumption power clusters

These differences help identify unexpected changes indicative of potential outages or disturbances. The anomaly score, indicated by the colored points, quantifies the magnitude of the deviation of each measurement from typical operating conditions. Higher anomaly scores, shown in darker colors, reflect larger deviations and thus more significant anomalies. The detected outages are further analyzed in Figure 5.29, where most outages are correctly identified. However, certain outages, highlighted within the pink boxes, remain undetected, and some locations are incorrectly classified as outages, despite no disconnection occurring in those areas. These discrepancies highlight areas where the algorithm's performance could be improved.

Figure 5.30 is included to understand the grid behavior during outages better. This figure

presents an example of the LV topology and the results of voltage distribution after power flow calculations for outages. Subfigure (a) corresponds to a detected outage, while subfigure (b) represents an undetected outage.

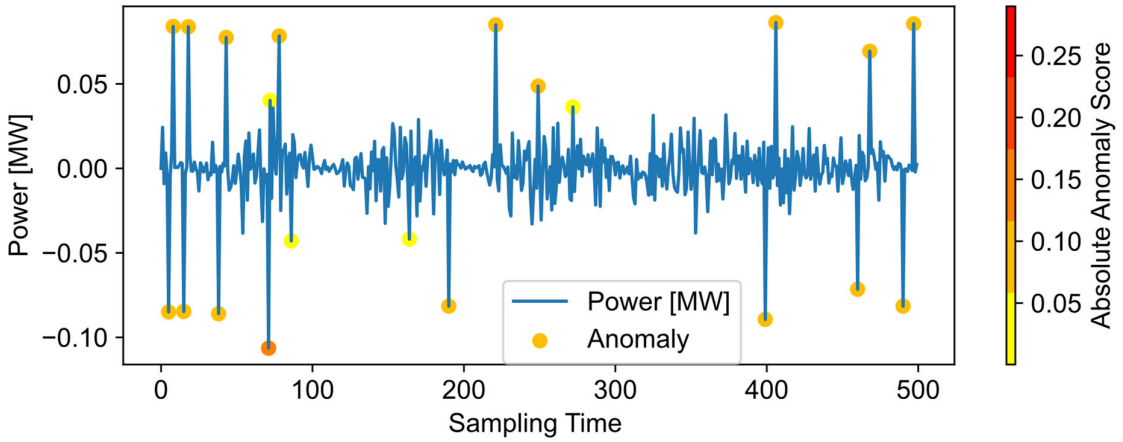


Figure 5.28: Outage detection on the differences in active power

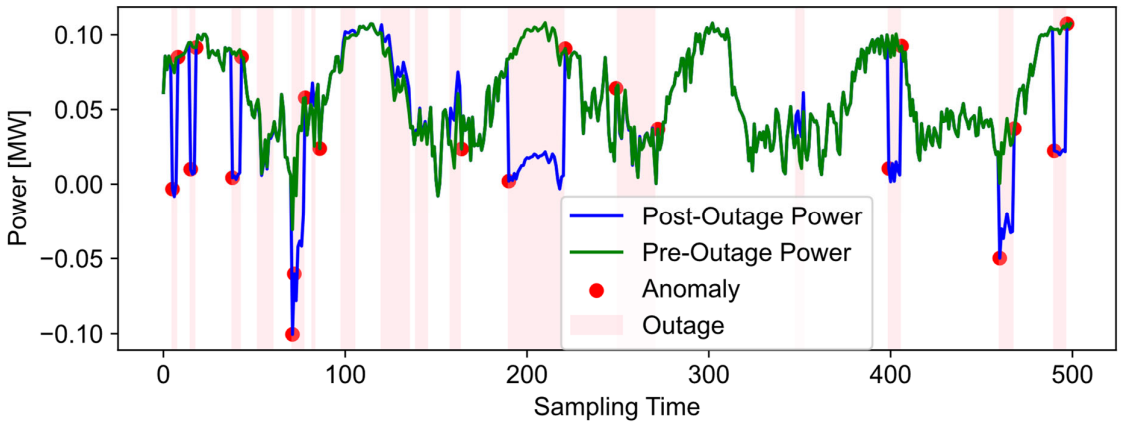
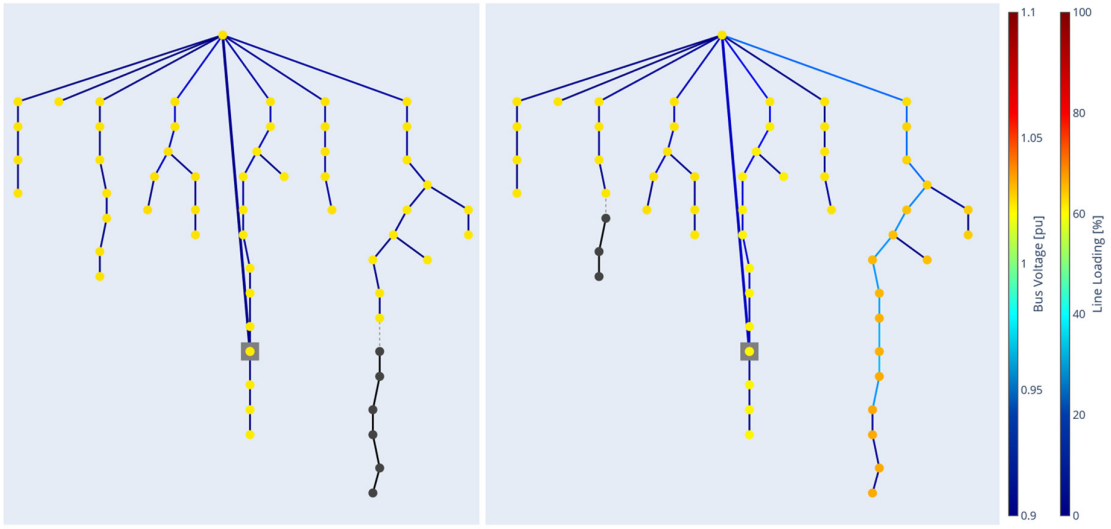


Figure 5.29: Results of the outage detection algorithm in active power

Figure 5.31 illustrates the relationship between the outage location, the number of disconnected nodes, and the corresponding detection results. Each bubble represents one outage event, where the x-axis denotes the distance from the busbar, the y-axis and bubble size indicate the number of disconnected nodes, and the color distinguishes detected (red) from undetected (gray) events.

The results show that the detection performance does not depend solely on the outage distance. While proximity to the busbar can influence the observable variations in power, the dominant factor is the magnitude of the disconnected load and the resulting change in power flow. Larger bubbles represent outages that disconnect more nodes and therefore cause greater power deviations, which are more easily detected by the algorithm. Conversely, smaller bubbles correspond to local or lightly loaded outages that produce weaker deviations and may remain undetected.



a) Grid topology and calculated voltage distribution for a detected outage

b) Grid topology and calculated voltage distribution for an undetected outage

Figure 5.30: Post-power flow outage topology and voltage profile.

Since the applied detection algorithm primarily evaluates power changes before and after an outage, events that lead to significant variations in power are detected with higher accuracy, regardless of their distance from the busbar. The lowest disconnected power that was successfully detected in this analysis was approximately 0.06 MW, indicating the algorithm's sensitivity threshold.

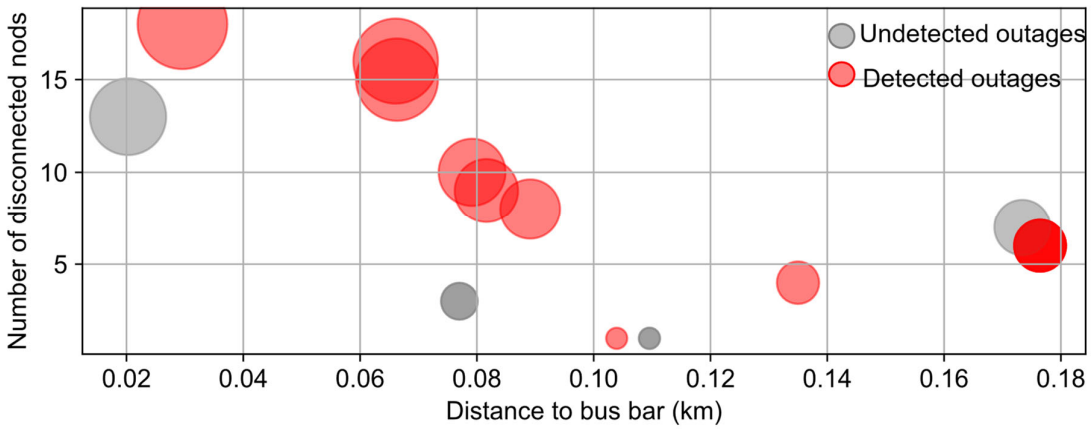


Figure 5.31: Correlation between busbar distance and node disconnection.

To improve the accuracy of detection, reactive power was incorporated as an additional parameter. While active power in LV grids is subject to significant short-term fluctuations due to normal load variability, reactive power typically exhibits smoother behavior and is less sensitive to routine load switching events. Consequently, deviations that appear only in active power are more likely attributable to normal load variations rather than to actual outages. In

contrast, outage events often result in abrupt and simultaneous changes in both active and reactive power due to the sudden disconnection of downstream loads. Based on this observation, a decision logic was applied in which anomalies detected in active power but not accompanied by corresponding changes in reactive power were filtered out as false positives. As shown by the orange circles in Figure 5.32, the inclusion of reactive power eliminated several of these false detections and significantly improved the overall detection accuracy.

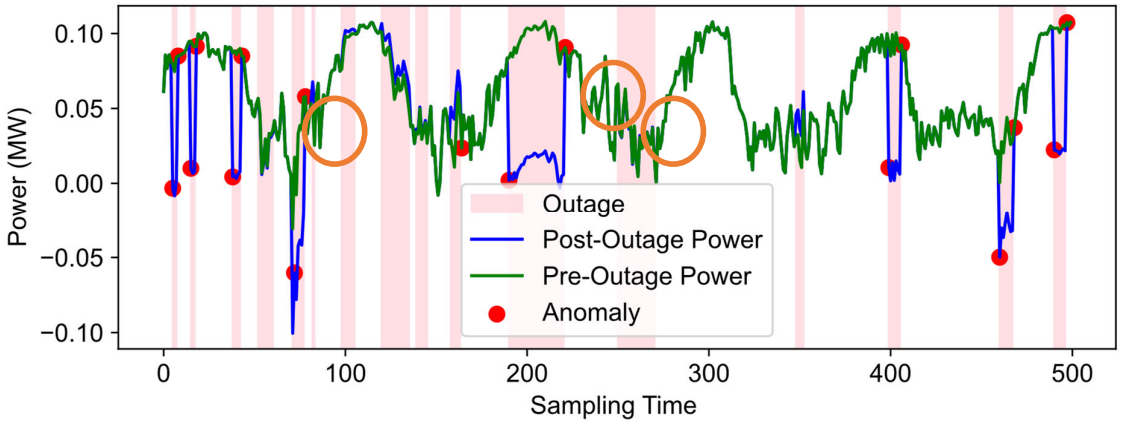


Figure 5.32: Impact of incorporating reactive power into the detection algorithm

Implementing CTs at key grid points is essential to improving outage detection. They provide precise current data for identifying even small outages and enhancing algorithm accuracy. Additionally, if smart meters are available on the grid, they can be used to pinpoint the location of outages accurately, further enhancing detection capabilities.

5.3 Summary and Research Focus

This chapter presents a detailed evaluation of multiple data-driven methods aimed at enhancing SA in LV grids. The work focuses on key challenges such as forecasting under high fluctuation, detecting contextual anomalies without labeled data, identifying blind spots in PV monitoring, and detecting outages using limited measurement infrastructure.

One of the main challenges in LV grids is the difficulty of precise forecasting under high variability caused by load changes and distributed PV generation. The PWCF method was introduced to address this issue by focusing on worst-case boundary prediction rather than exact point forecasting. While PWCF demonstrates strong performance, it faces limitations due to the lack of real-time data from interconnected grid components. This missing information, such as transformer tap changes or reactive power support from nearby feeders, affects voltage and current behaviors, thereby influencing forecast reliability.

To further enhance SA, several anomaly detection methods were employed. A major challenge in this context is the absence of clearly labeled anomalies, making supervised learning impractical. Additionally, the anomalies of interest are often contextual rather than point-based, that is, they only appear anormal when considering environmental or temporal context. To

overcome this, hybrid approaches combining STL method with ML methods, as well as LSTM Autoencoders, were implemented. These methods effectively decomposed normal patterns and identified meaningful deviations without requiring labeled data.

Despite the strengths of ML-based methods, statistical techniques remain valuable, especially in scenarios where interpretability, low computational cost, or limited data are critical. Methods such as CUSUM and PDF-based seasonality tracking can detect gradual shifts and recurring pattern changes, making them useful for grid operation and planning.

Another layer of SA improvement is blind spot detection, particularly the identification of unregistered PV installations. The clustering-based detection strategy highlighted in this chapter successfully revealed potential PV generation not captured by system records. The challenge here lies in accurately estimating consumption by filtering out modeled PV generation, especially in the presence of single-phase or battery-based systems.

Finally, the chapter explores outage detection using only busbar-level measurements. The primary challenge is limited observability: without detailed nodal data, small-scale outages or those occurring far from the busbar can go undetected. Initially, detection relied solely on active power differences, which led to misclassifications due to regular load changes. Incorporating reactive power as a complementary feature significantly improved detection accuracy by stabilizing the detection signal. Future enhancements, such as installing current transformers (CTs) or utilizing smart meter feedback, could further refine outage localization and reduce false positives.

Together, these methods provide a layered approach to SA in LV grids, combining forecasting, anomaly detection, blind spot identification, and outage detection to address current limitations and anticipate future operational needs.

6 Discussion, Conclusion, and Future Work

This chapter concludes the thesis by discussing the broader implications of the presented research, synthesizing the key outcomes, and reflecting on the methodological contributions. The aim is to critically evaluate the effectiveness of the approaches developed and their potential application in assisting decision-making in the LV grids. In doing so, this chapter emphasizes the relevance of data-driven and Artificial Intelligence (AI)-supported techniques in addressing emerging challenges in modern power distribution systems.

Moreover, the chapter highlights the limitations encountered during the study and outlines areas where further investigation is needed. In addition to summarizing the core findings, it introduces future research directions that could extend the applicability and robustness of the proposed solutions. By doing so, the chapter aims to provide a comprehensive perspective on how the work contributes to the ongoing advancement of Situational Awareness (SA) and intelligent decision-making in LV grid environments.

6.1 Discussion and Conclusion

This thesis represents methods to improve SA in distribution grids, with a targeted study of LV grids. This work identifies and addresses several challenges, particularly the lack of measurement infrastructure, which significantly limits grid observation. This lack of understanding and analysis of grid behavior often hinders and leads to suboptimal decisions and increased susceptibility to anomalies, disrupting operations and complicating management under extreme conditions. The complexity of LV grids is constantly increasing, mainly due to the integration of renewable energy sources and the changing demands of new electrical loads.

In response to these challenges, Digital Process Twin (DPT) is a solution to enhance the coherence and reliability of the grid's operational data. Through DPT, Distribution System Operators (DSOs) can achieve a more unified and accurate view of the grid's status. This system ensures that changes within the grid are promptly considered, and data integrity is upheld through rigorous processing checks before storage, thus guaranteeing the data’s unification and validity.

Through the deployment of DPT, data and AI-driven methods have been introduced to provide assistive signals through the knowledge extraction component of DPT, considering both forecasting and anomaly detection. Forecasting in the LV grid faces challenges due to the high fluctuation of data, which makes accurate short-term forecasting particularly difficult. To address this issue and assist in control and decision-making within DSOs, the Pseudo-Worst Case Forecast (PWCF) method has been introduced. This tackles the challenges of high data fluctuation and the lack of accurate grid topology by employing a topology-independent

method. However, this method needs further improvement, which could be achieved by considering external data from the medium voltage grids, such as control commands.

This thesis explored different methods for anomaly detection from the consumer side in LV grids, focusing particularly on data-driven methods like AI methods, including Machine Learning (ML) methods and deep learning. These AI methods are renowned for their adaptability and automation, crucial for handling the complexities of LV grids. However, their effectiveness depends on several factors such as the quality and relevance of the data used, the specific goals of the application, the required accuracy, the complexity of the tasks, and the computational demands.

One significant challenge identified in this research is that AI models struggle when there are major changes in grid patterns or the structure of the grid itself. This issue arises because the training data the models learn from no longer matches new conditions, underscoring the need for models that can continuously adapt and learn on their own without constant manual updates. In addition, AI methods have their "black box" nature, where the internal workings and the training process are not transparent. If the input data includes anomalies or errors, the accuracy of the results can decrease significantly.

In this thesis, unsupervised learning methods for anomaly detection were considered. The reasons supervised learning methods were not pursued include their reliance on carefully labeled data to train the models. Labeling involves identifying and categorizing data points such as changes in energy use, new equipment installations, or temperature shifts. However, labeling is challenging because it heavily depends on the operators' ability to correctly recognize and classify anomalies. Additionally, when new types of anomalies occur, the system needs to be updated with new labels. In contrast, unsupervised learning methods do not require predefined labels and can automatically detect and adjust to new patterns. This capability is especially useful in LV grids, where changes can be unpredictable and frequent, making unsupervised methods more effective for continuous monitoring and analysis.

In conclusion, this thesis demonstrates that the selection of methodologies in data-driven approaches is not a unique solution but depends heavily on the available data and specific use cases. Throughout this study, various methods were explored, and new ideas for data-driven approaches were introduced. However, these methods are not universally optimal for all types of data and use cases. Each method has its strengths and limitations, emphasizing the need for a tailored approach based on the specific requirements and conditions of the LV grids.

6.2 Future Work

Future work could focus on analyzing the impact of changes in the training data pattern. Specifically, if there are major shifts in the pattern, it is important to investigate how the accuracy of machine learning models can be maintained. If accuracy is negatively affected,

one possible solution could be to reduce the influence of outdated or less relevant historical data that no longer reflects the current state of the system.

Another direction for future research is to explore physics-informed machine learning methods that take into account the grid topology. These methods could adapt to changes in the grid structure and provide more robust predictions. Additionally, using data from different types of grids and evaluating how asset and component changes affect anomaly detection is essential. Since the definition of an anomaly often depends on the contextual state of the grid, changes in topology, such as switch operations, could alter what is considered an anomaly.

Another future work is the optimal management of flexibility services in LV grids by leveraging signals derived from knowledge extraction methods. Signals such as anomalies detected through DPT or outputs from PWCF can serve as valuable indicators for grid behavior. These signals can be applied in various use cases, including predictive maintenance, detection of operational risks, and flexibility assessment. For instance, a reduction in grid flexibility due to the integration of new loads can be identified early through these signals, enabling timely intervention.

List of abbreviations

Abbreviation:	Term:
AI	Artificial Intelligence
AISOP	AI-assisted Support for Operational Planning
CT	Current Transformer
CUSUM	Cumulative Sum
DPT	Digital Process Twin
DSO	Distribution System Operators
DT	Digital Twin
DTiPS	Digital Twin in Power Systems
DWD	Deutscher Wetterdienst
ESD	Extreme Studentized Deviate
FBWM	Frequent Behavior Weighted Mean
iONS	inteligente Ortsnetzstationnen
IoT	Internet of Things
IQR	Interquartile Range
KDE	Kernel Density Estimation
KLD	Kullback-Leibler Divergence
LoRaWAN	Long Range Wide-Area Networks
LPWAN	Low Power Wide Area Network
LSTM	Long Short-Term Memory
LV	Low Voltage
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NN	Neural Networks
OT	Operational Technology
PDF	Probability Density Function
PMF	Probability Mass Function
PV	Photovoltaic
PWCF	Pseudo-Worst Case Forecasting
RES	Renewable Energy Sources
SA	Situational Awareness
SCADA	Supervisory Control and Data Acquisition
SmartAPO	Smart Autonomer Prädiktiver Ortsnetzregle
SSoT	Single Source of Truth
STL	Seasonal and Trend Decomposition using Loess

List of symbols

Variable:	Meaning:	Base unit:
$\boldsymbol{v}_{base}^{i}$	Daily voltage pattern at node i	[Volt]
$\boldsymbol{v}_{l,border}^{i}$	Lower voltage border at node i	[Volt]
$\boldsymbol{v}_{u,border}^{i}$	Upper voltage border at node i	[Volt]
$\boldsymbol{v}_{l,border,adj}^{i}$	Adjusted the lower voltage border at node i	[Volt]
$\boldsymbol{v}_{u,border,adj}^{i}$	Adjusted upper voltage border at node i	[Volt]
V_{his}^{i}	Historical voltage matrix for each node	[Volt]
$\boldsymbol{v}_{D,H}^{i}$	Sample vector for day D and hours H at node i	[Volt]
$\overline{\mathrm{V}}(t_0)$	Smoothed value	[Volt]
$\overline{\boldsymbol{v}}_{base}^{i}$	Smoothed daily voltage pattern at node i	[Volt]
$\mathbf{V}_{\mathbf{s}}^{i}$	Sorted historicalvoltage matrix	[Volt]
$\boldsymbol{v}_{\max}^{\mathrm{i}}$	Vector value of maximum voltages	[Volt]
$\boldsymbol{v}_{\min}^{\mathrm{i}}$	Vector value of minimum voltages	[Volt]
$V_{lt,adj}^{i}$	Long-term adjusted value	[Volt]
$\mathrm{V}_{st,adj}^{i}$	Short-term adjusted value	[Volt]
$\boldsymbol{v}_{u,fc}$	Upper border vector for voltage	[Volt]
$\boldsymbol{v}_{l,fc}$	Lower border vector for voltage	[Volt]
$\boldsymbol{v}_{in,t_0}$	Input vector of historical voltage data at time t_0	[Volt]
$\boldsymbol{v}_{out,t_0+\Delta t}$	Output vector of historical voltage data at time $t_0 + \Delta t$	[Volt]
$\boldsymbol{v}_{in\_clustered,t_0}$	Input vector of clustered historical voltage data at time t_0	[Volt]
$\boldsymbol{v}_{out,t_0+\Delta t}$	Output vector of clustered historical voltage data at time $t_0 + \Delta t$	[Volt]
V_{i,t_0}^{C}	Clusterd data of node i at time t_0	[Volt]
$\mathrm{V}_{i,t_0}^{diff}$	Deviations between the actual observed data and the basic pattern data	[Volt]

Variable:	Meaning:	Base unit:
$\mathrm{V}_{diff,max}^{i,t_0}$	Difference from the upper historical maximum	[Volt]
$\mathrm{V}_{diff,min}^{i,t_0}$	Difference from the lower historical minimum	[Volt]
$d_{\mathrm{worst,max}}^{i,t_0+\Delta t}, d_{\mathrm{worst,min}}^{i,t_0+\Delta t}$	Maximum and minimum deviation scaling	
$\alpha_{\max}, \alpha_{\min}$	Scaling factors for deviation adjustment	
μ_w, σ_w	Mean and standard deviation within the window *w*	
y	Time series data or target variable (e.g., net power)	[Watt]
$\mathbf{y}_{\mathrm{trend}}, \mathbf{y}_{\mathrm{s}}, \mathbf{y}_{\mathrm{ress}}$	Trend, seasonal, and residual components	[Watt]
$\mathbf{c}_{\mathrm{HPV}}, \mathbf{c}_{\mathrm{LPV}}$	Clusters for high and low PV power potential	-
$\mathbf{Z}_{\mathrm{res}}$	standardized scale	-
λ_{k}	ESD factor	-
X	Time series input	[Watt]
$T_i, S_{p,i}, R_i$	Trend, seasonal, and residual components	[Watt]
$\mathbf{F}_{\mathrm{T}}$	Feature matrix of the target window (for correlation analysis)	-
$\mathbf{F}_{\mathrm{W_i}}$	Feature matrix of window *i* (used to compute similarity to FT)	-
$\mathbf{y}_{\mathrm{Top_k}}$	Data from top *k* windows with highest correlation	[Watt]
$\mathbf{y}_{\mathrm{W_T}}$	Data from the selected window for anomaly detection	[Watt]
DT_i	Detrended data at time *i*	[Watt]
$DT_{p,i}$	Detrended data after removing *p*-th seasonality	[Watt]
C_k^{DT}	Cyclic subseries of detrended data used for estimating seasonality	[Watt]
M_t	Frequent Behavior Weighted Mean (FBWM), computed from clustered historical data	[Watt]
$\overline{y}_{t,k}$	Mean of data points in cluster k at time t	[Watt]
C_k	Number of observations in cluster k	-
$\mathrm{R_t}$	Normalized ratio	-

Variable:	Meaning:	Base unit:
Z_t	Z-standardized value of y_t	-
S_T	Cumulative sum of Z_t values up to time t	-
$Prob_t$	Probability of a change point at time t	-
$\Phi(\widetilde{s_t})$	CDF of standard normal at $\widetilde{S_T}$	-
$\hat{f}(y)$	Estimated probability density function via KDE	-
$K(y)$	Kernel function (e.g., Gaussian) used in KDE	-
$KLD(P \parallel Q)$	Kullback-Leibler Divergence between distributions P and Q	-
$P_{\max}$, $P_{\min}$	Max/Min power values within cluster c	[Watt]
P_t	Measured power at time t	[Watt]
E_{max}, E_{min}	Deviation of $P_t$from upper/lower bounds (anomaly scores)	[Watt]
ΔP_t , ΔPQ_t	Difference in active/reactive power between t and $t-1$	[Watt]

Appendix

A. CUSUM Method

The p-value is the probability under the null hypothesis of obtaining a real-valued test statistic at least as extreme as the one obtained.

$$\begin{aligned}\mathrm{P}\left(\left|\widetilde{S}_t\right| \geq |\widetilde{s}_t|\right) &= 1 - \mathrm{P}\left(-|\widetilde{s}_t| \leq \widetilde{S}_t \leq |\widetilde{s}_t|\right) = 1 - \left[\mathrm{P}\left(\widetilde{S}_t \leq |\widetilde{s}_t|\right) - \mathrm{P}\left(\widetilde{S}_t \leq -|\widetilde{s}_t|\right)\right] \\ &= 1 - \left[\mathrm{P}\left(\widetilde{S}_t \leq |\widetilde{s}_t|\right) - \left(1 - \mathrm{P}\left(\widetilde{S}_t \leq |\widetilde{s}_t|\right)\right)\right] = 2 \cdot \left(1 - \mathrm{P}\left(\widetilde{S}_t \leq |\widetilde{s}_t|\right)\right) \\ &= 2 \cdot \left(1 - \Phi(|\widetilde{s}_t|)\right)\end{aligned}$$

B. iONS Stations

"iONS" refers to "intelligent Ortsnetzstationen", local network substations, key nodes in the electricity grid, that facilitate the distribution of energy to local LV networks and are increasingly being digitized to optimize grid management. These substations are crucial for the advancement of energy transition as they support load management functions within the networks, thereby enhancing the integration of solar, wind, and heat pump systems. The digitization of these nodes allows for more precise monitoring and control of energy flow, leading to improved grid stability and efficiency. Moreover, the use of smart measurement and control technologies enables early detection and resolution of bottlenecks, further accelerating the integration of renewable energy sources. Thus, local network substations play a central role in ensuring a reliable and sustainable energy supply by adaptively responding to the dynamic requirements of a modern energy grid. Important devices in the system include the feeder condition monitor, which keeps track of electrical conditions and identifies faults like short circuits or ground faults.

Another key device is the power quality recorder, which logs data and checks for any issues with the electricity quality, ensuring the power supplied meets required standards. The system also features a smart grid Remote Terminal Unit (RTU) with programmable controls and a smart grid unit, which handles communications within the grid.

Benefits of smart local network substations [159]:

- Monitoring and ensuring energy quality: These substations help maintain high standards of power quality.
- Managing overloads: They effectively handle situations where the system is at risk of becoming overloaded.
- Minimizing downtime: By reducing the duration of interruptions, they decrease the loss of network fees.
- Optimizing network expansion: Smart substations facilitate more efficient upgrades and

expansions of the power network.

- Substation monitoring: Continuous monitoring of the substations improves overall management and responsiveness.

While expanding the network can provide the necessary capacity for these energy sources, only intelligent solutions can effectively manage changing directions of energy flow, load fluctuations, and maintain voltage levels.

Figure B. 1 visual represents the placement and role of iONS within a power grid, as highlighted by the blue circles.

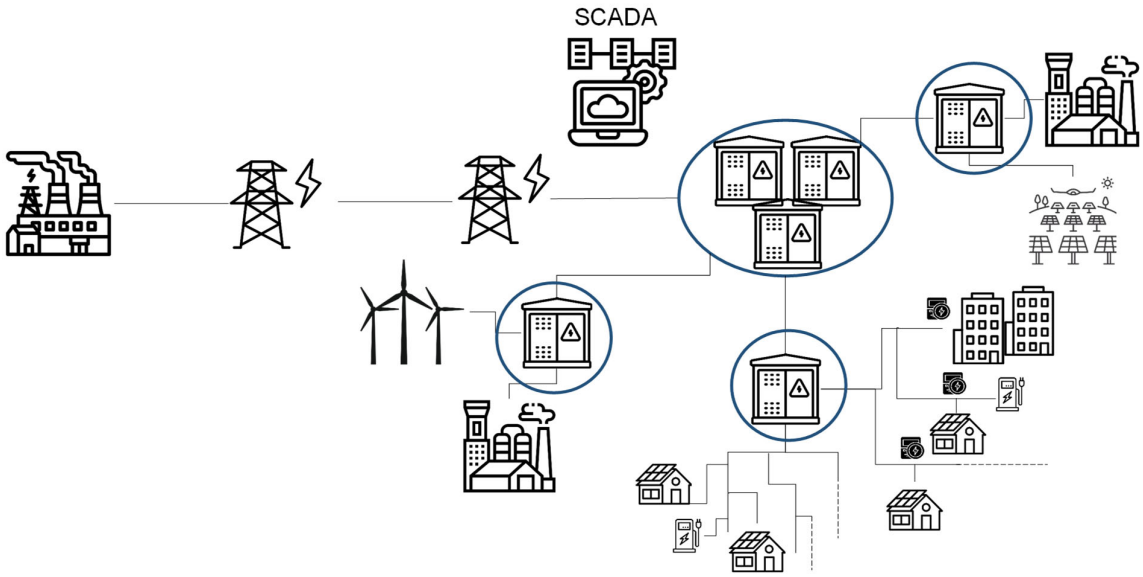


Figure B. 1: Integration of intelligent iONS in the power grid

C. SMAPO Data

Figure C. 1 shows the grid layout for the SMAPO data, which includes 93 nodes.

D. Methods for Detecting Rural and Urban Station Types

In this thesis, the anonymity of station types presented a challenge for simulating results and finding correlations between stations. Identifying and distinguishing between urban and rural LV stations in Germany was useful to facilitate this analysis. Here, two methods are introduced for detecting the types of stations based on various characteristics.

Some important features that could be helpful to distinguish the types of stations are as follows:

- Location: Urban stations are often found in densely populated residential neighborhoods, commercial areas, or industrial zones. Conversely, rural stations are typically located in areas with low population density, such as farmland or small villages.

- Size: Urban stations usually cover a smaller area but serve a higher concentration of customers, making them smaller in size. Rural stations may be larger due to the need to serve a more extensive area with fewer customers.

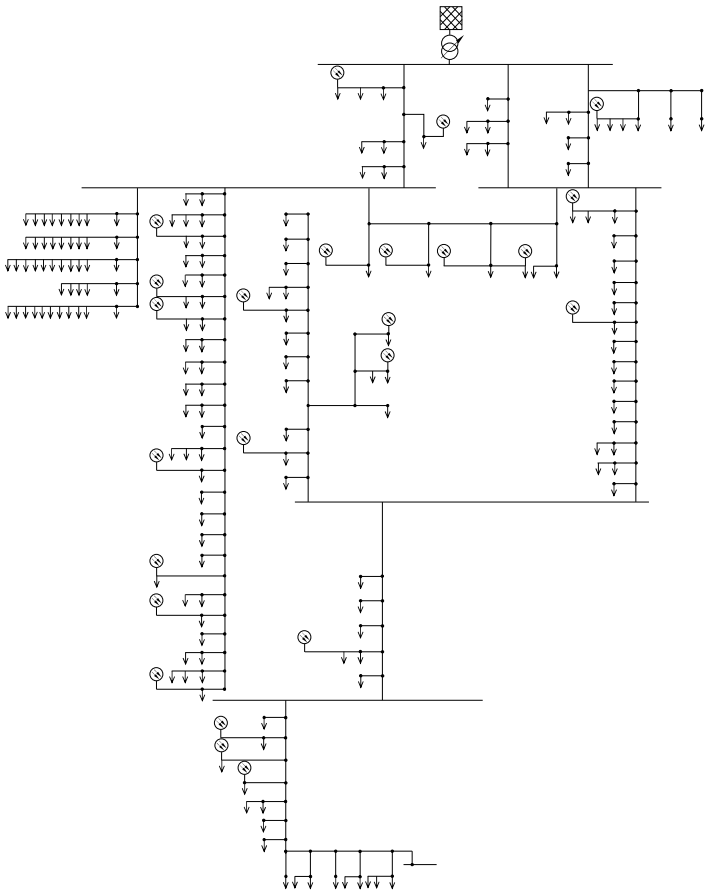


Figure C. 1: SMAPO topology of the grid

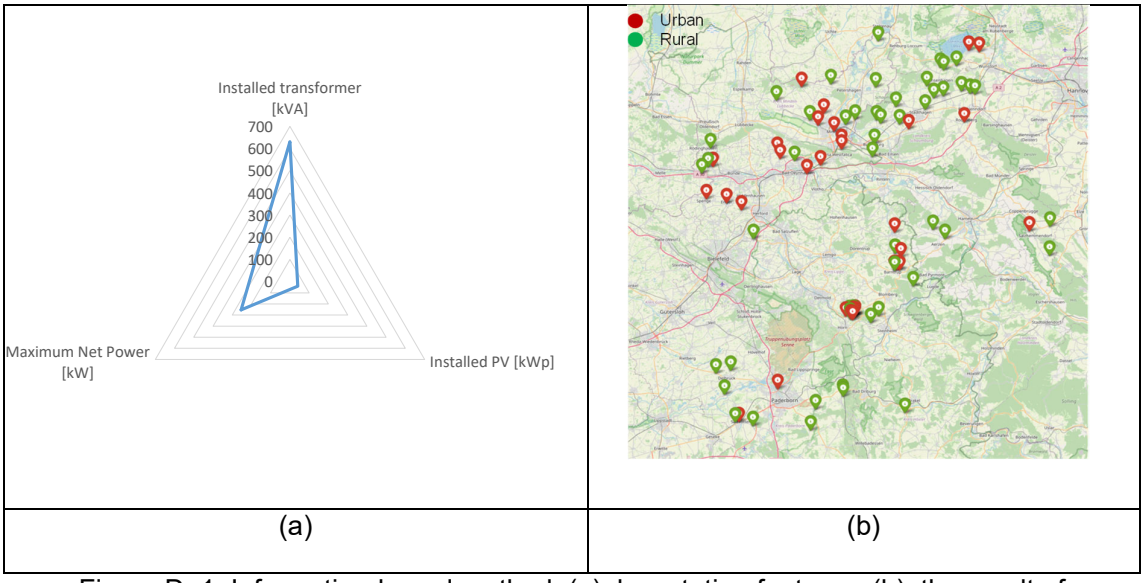
- Infrastructure: Urban stations may feature more advanced infrastructure, such as underground cabling, to meet the needs of a densely populated area. Rural stations are more likely to rely on above-ground equipment and older infrastructure.
- Electricity demand: Urban stations generally face higher electricity demand due to the greater number of customers. In contrast, rural stations have a lower demand, corresponding with the lower population density of the areas they serve.
- Customer data: The number of customers served by a station can indicate its location. Urban stations typically serve more customers, while rural stations serve fewer.

Identifying whether a station is urban or rural provides insights into its surrounding area's population density, electrical demand, and infrastructure availability. The classification includes urban, semi-urban, rural, and semi-rural categories. Proximity rules apply where stations closest to urban or semi-urban areas are labeled urban, and similarly for rural.

- Information-based method: This approach compares station characteristics to determine their classification. For example, a station with a rate of installed PV that exceeds a specific threshold and substantial PV capacity might be categorized as rural. Conversely, if a station's installed transformer capacity significantly exceeds its PV installations, it will likely be classified as urban.

Figure D. 1 (a) presents spider charts with detailed information that supports an information-based method for distinguishing station types. Figure D. 1 (b) displays the results of this method on a map, where urban stations are marked in red and rural stations are depicted in green.

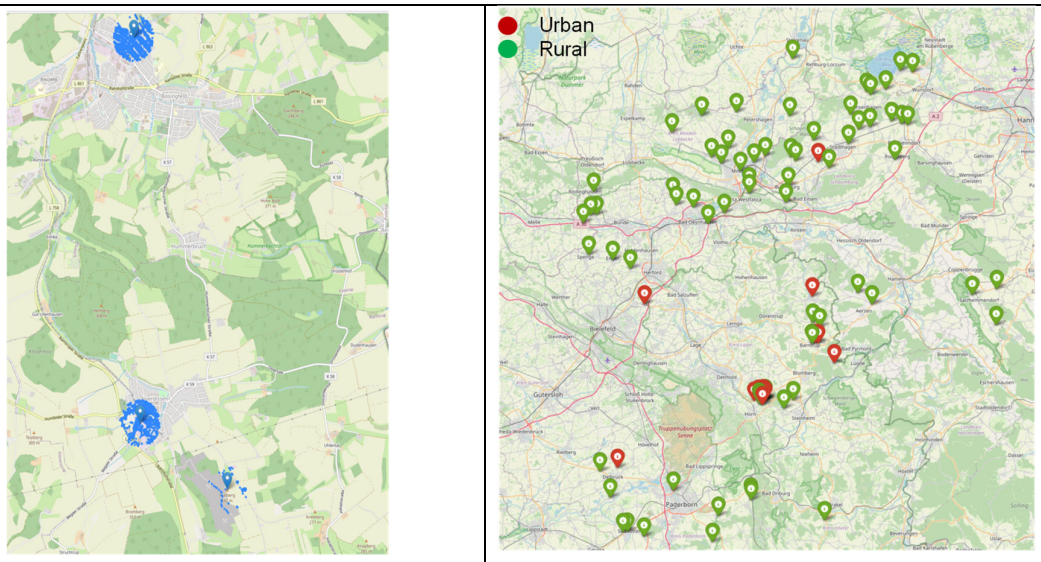
- **Location-based method:** This method considers the geographical placement of the stations to categorize them. It leverages mapping and spatial data to infer urban or rural status based on relative isolation or integration within populated areas. Additionally, using the station's address, we can estimate the number of houses in its vicinity. For this estimation, a radius of 3 km around the station's address is considered. However, it's important to note that, due to the predominantly radial topology of most LV grids, this method may not accurately distinguish the specific area associated with the station. Nevertheless, it does provide insight into whether the station is located in a densely populated area



(a) (b)

Figure D. 1: Information-based method, (a): key station features, (b): the result of distinguishing the type of stations

Figure D. 2 is an illustration of the Location-Based Method. This figure shows how the location-based method operates, with dots representing the corners of each household. A higher concentration of dots indicates a densely populated area, typically characterizing an urban station.



(a) (b)

Figure D. 2: Location-based method. (a) blue dots represent the density of surrounding stations; (b) the result of distinguishing the type of stations

These two methods differ in distinguishing station types. For instance, one method may classify a station as rural, whereas the other identifies it as urban. To refine this approach, it is recommended that stations be considered individually. This involves examining their energy behavior over a year or analyzing their location on the map to determine the types of consumers nearby. For this thesis, the stations selected for simulation consistently showed the same results with both methods.

E. Method for Selecting High-Fluctuation Case Study Stations

The flowchart (Figure D. 3) outlines a method for analyzing and categorizing stations to select interesting case studies for simulation. This involves clustering stations based on similarities in their net power time-series data. Clusters with significant fluctuations are identified and further classified into different types, such as urban and rural. From these categories, stations that exhibit a strong correlation with weather information are selected, resulting in a subset of "selected stations" that represent high fluctuation behavior and diverse types (urban and rural). These selected stations are most impacted by external factors like weather and are ideal for detailed analysis and simulation.

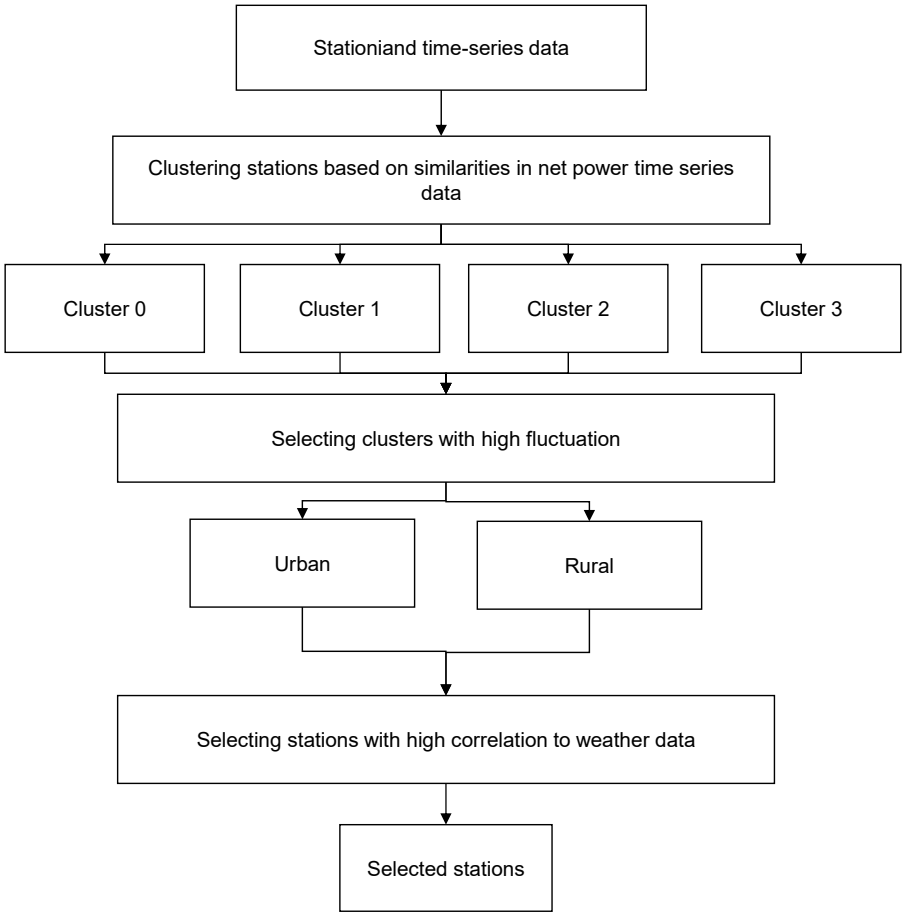


Figure E. 1: Flowchart for selecting Stations with high fluctuation behavior for the case study

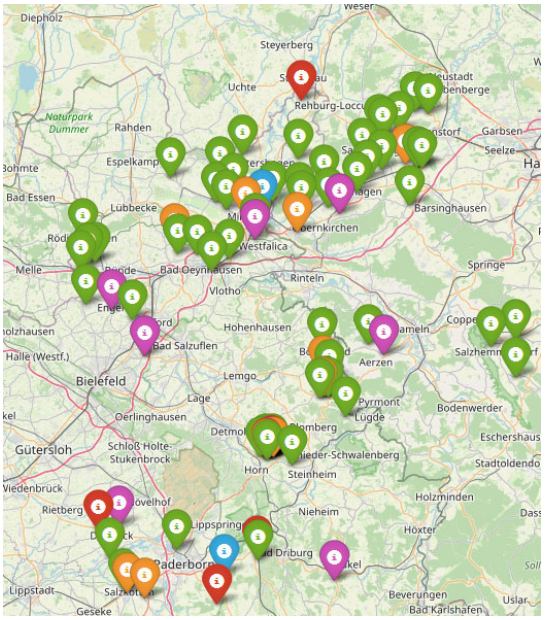
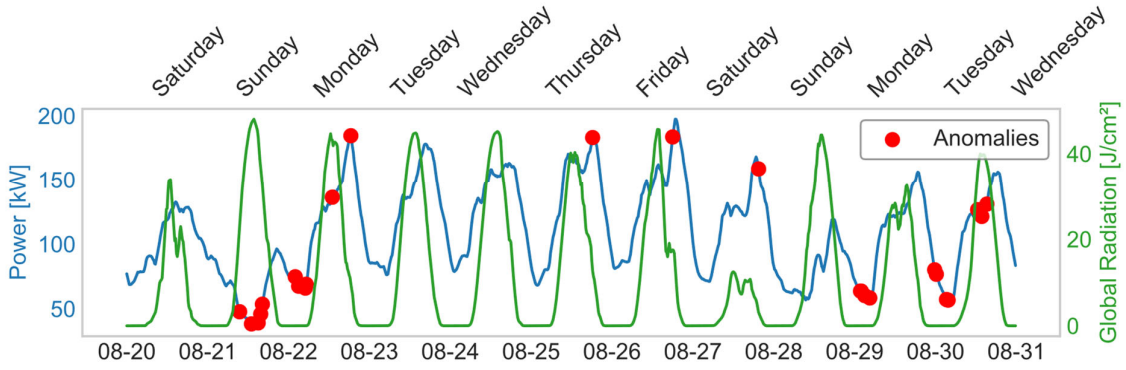


Figure E. 2: Clustered stations based on similarity in net power time-series behavior

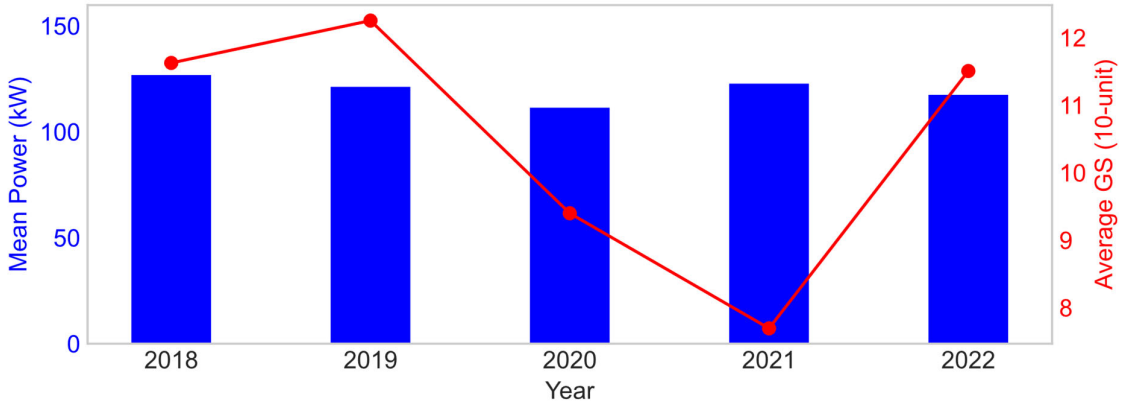
F. Simulation Results of AI-based Anomaly Detection Methods

Cluster-Enhanced Feature Regression

Figure F. 1 (a) presents the results of the method cluster-enhanced feature regression, highlighting anomalies: (a) A significant data drop on 21.08, due to extra photovoltaic installations, (b) Decreased nighttime power consumption on the nights of 22, 29, and 30 August, (c) An increase in daytime power usage, such as 27.08, considered anomaly due to reduced photovoltaic generation; (d) Minor pattern shifts and inconsistencies in training data or weather information also contribute to the anomalies. Figure F. 1 (b) shows that, compared to 2018, with the same mean global sun radiation, the energy usage in 2022 is less. This reduction could be attributed to the additional photovoltaic installations, which generate more power.



a) Net power data anomalies and global sun radiation



b) Mean power values and mean solar radiation over the same period across all years

Figure F. 1: The results of the cluster-enhanced feature regression method

Hybrid Extreme Studentized Deviate

Figure F. 2 showcases the results of the method, hybrid extreme studentized deviate, which successfully pinpointed significant reductions in energy usage resulting from enhanced PV generation. This method is more sensitive than previous ones because it incorporates adjustments for residual data, leading to the identification of a greater number of anomaly points.

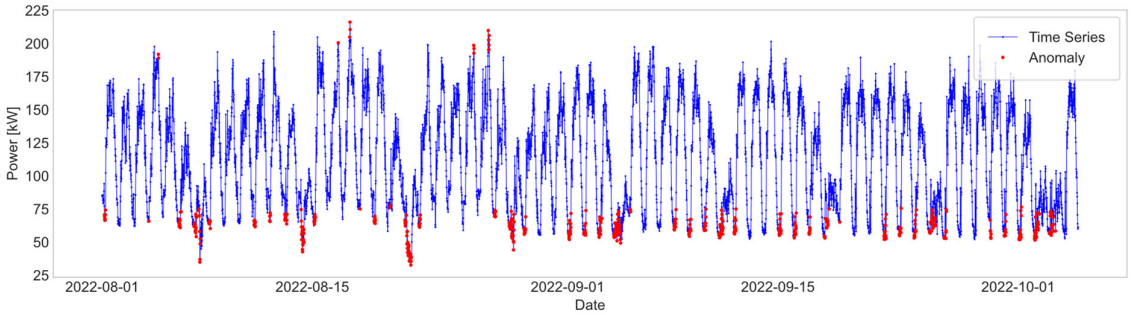


Figure F. 2: The results of the analysis of the hybrid ESD

Statistical Windowing with Isolation Forest

Figure F. 3 shows the results of the method of statistical windowing with isolation Forest, which incorporates the use of a rolling average to enhance the detection process. This approach effectively filters out minor fluctuations, ensuring that only significant deviations are classified as anomalies.

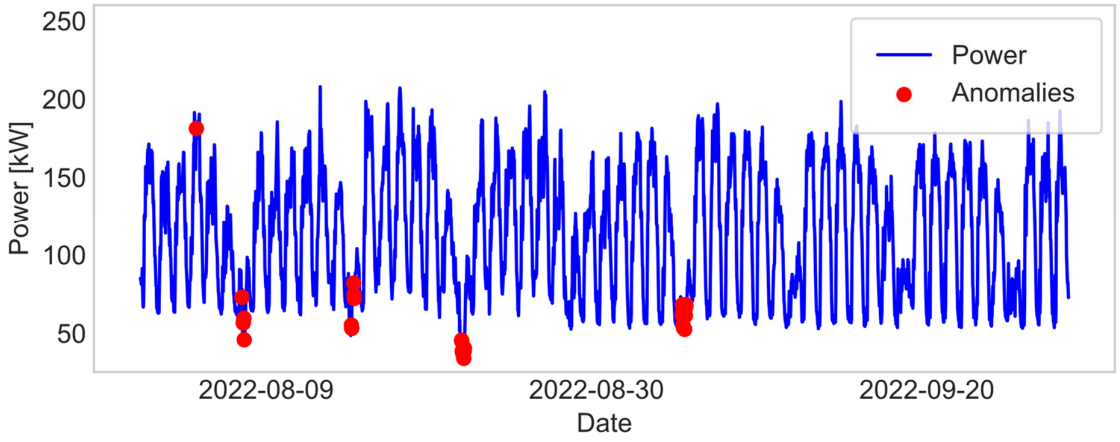
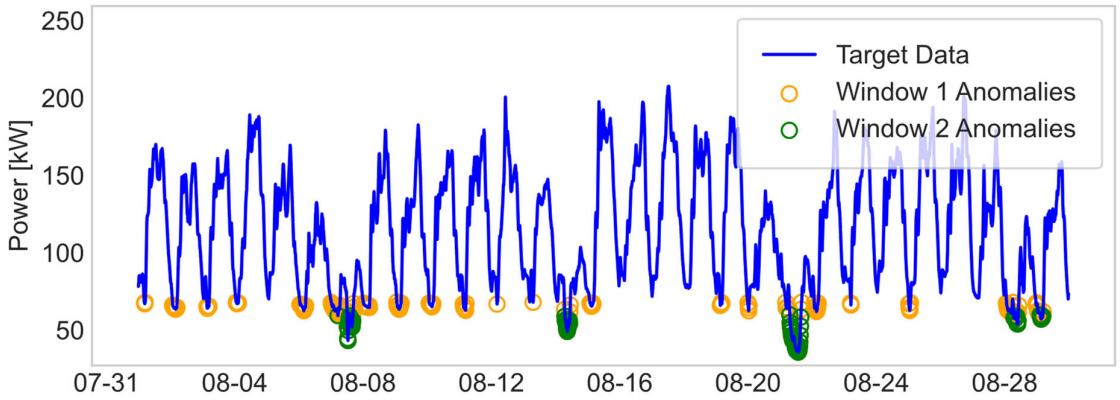


Figure F. 3: The results of the statistical windowing with isolation Forest

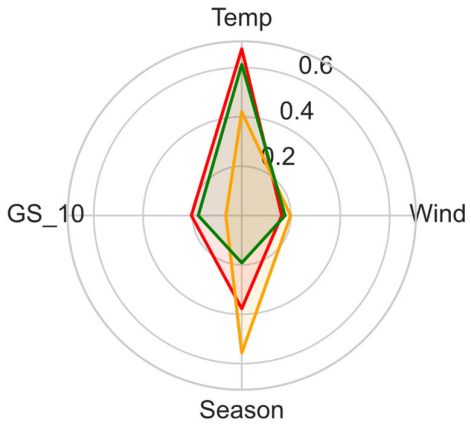
Correlative Isolation Forest

The target period for this method was from 01.07- 30.09.2022. Figure F. 4 (a) highlights anomalies in two historical windows with high correlation to the target period: Window 1 (01.01-16.04.2017) and Window 2 (01.04-21.06.2017). Figure F. 4 (b) emphasizes window 2's strong correlation with the target. In Window 1, minimum energy usage was higher than in the target period. In Window 2, decreases in energy due to extra PV installations were flagged as

anomalies. Anomalies in maximum values were linked to changes in usage patterns, identifying PV installations under similar conditions.



a) Illustrate anomalies within two historical windows



b) Average feature comparison of target data vs correlated historical data

Figure F. 4: The results of the correlative isolation Forest method

References

[1] J. W. Ellis, Fundamentals of homeland security: An operations perspective: Charles C Thomas Publisher, 2014.

[2] Y. Zhang, C. Qin, A. K. Srivastava, C. Jin, and R. K. Sharma, "Data-driven day-ahead PV estimation using autoencoder-LSTM and persistence model," IEEE Transactions on Industry Applications, vol. 56, no. 6, pp. 7185–7192, 2020.

[3] C. Qin, J. Greig-Prine, Z. Nie, P. Banerjee, and A. K. Srivastava, "Remote PMU testing using low-cost FPGA platform and PPA following IEEE TSS," IEEE Industry Applications Society Annual Meeting, 2019.

[4] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," Renewable and sustainable energy reviews, Elsevier journal, vol. 56, pp. 215–225, 2016.

[5] M. Ferdowsi, A. Benigni, A. Löwen, B. Zargar, A. Monti, and F. Ponci, "A scalable data-driven monitoring approach for distribution systems," IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 5, pp. 1292–1305, 2015.

[6] H. Ren, Z. Hou, X. Ke, Q. Huang, and Y. Makatov, "Analysis of weather and climate extremes impact on power system outage," IEEE Power & Energy Society General Meeting (PESGM), 2021.

[7] J. McCarthy et al., "Situational awareness for electric utilities," National Institute of Standards and Technology Special Publication, 7B, 1800. [Online]. Available: https://doi.org/10.6028/NIST.SP.1800-7

[8] C. Chen, J. Wang, and D. Ton, "Modernizing distribution system restoration to achieve grid resiliency against extreme weather events: An integrated solution," Proceedings of the IEEE, vol. 105, no. 7, pp. 1267–1288, 2017.

[9] Y. Zhang, "Model-based and data-driven situational awareness for distribution system monitoring and control," Ph.D. dissertation, Dept. of Electrical Engineering, Southern Methodist University, 2020.

[10] M. A. A. Al-Jaafreh and G. Mokryani, "Planning and operation of LV distribution networks: a comprehensive review," IET Energy Systems Integration, vol. 1, no. 3, pp. 133–146, 2019.

[11] D. Forstmann, "Messen, Regeln, In allen Netzen.: Moderne Power-Quality Lösungen mit Abgangsmesstechnik & Erdschlussanzeiger als Digitalisierungseinheit in Ihrer Ortsnetzstation," 2023. [Online]. Available: https://www.driescher-wegberg.de/hm2023/Vortrag_A.Eberle.pdf

[12] R. Singh, E. Manitsas, B. C. Pal, and G. Strbac, "A recursive Bayesian approach for identification of network configuration changes in distribution system state estimation," IEEE Transactions on power systems, vol. 25, no. 3, pp. 1329–1336, 2010.

[13] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying topology of low voltage distribution networks based on smart meter data," IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 5113–5122, 2017.

[14] R. van Dinter, B. Tekinerdogan, and C. Catal, "Predictive maintenance using digital twins: A systematic literature review," Information and Software Technology, Elsevier, vol. 151, p. 107008, 2022.

[15] D. M. Hawkins, "Identification of Outliers," London, U.K.: Chapman & Hall, Springer, vol. 11, 1980.

[16] K. Moloi, M. Ntombela, Thapelo. C. Mosetlhe, Temitope. R. Ayodele, and Adedayo. A. Yusuff, "Feature extraction based technique for fault classification in power distribution system," IEEE PES/IAS PowerAfrica, 2021.

[17] W. H. Wellßow et al., "Advantages of digitalization in grid restoration," at-Automatisierungstechnik, De Gruyter Oldenbourg, vol. 68, no. 9, pp. 790–803, 2020.

[18] M. Panteli and D. S. Kirschen, "Situation awareness in power systems: Theory, challenges and applications," Electric Power Systems Research, Elsevier, vol. 122, pp. 140–151, 2015.

[19] G. Hofbauer, Z. Ting, D. Maier, and L. Zhi, "Digital Twin Application for Smart Product Management and Production Engineering and Deployment in China," Technischen Hochschule Ingolstadt, 2024.

[20] A. Aghazadeh Ardebili et al., "Enhancing resilience in complex energy systems through real-time anomaly detection: a systematic literature review," Energy informatics, Springer, vol. 7, no. 1, p. 96, 2024.

[21] X. He, Q. Ai, R. C. Qiu, and D. Zhang, "Preliminary exploration on digital twin for power systems: Challenges, framework, and applications," IEEE Transactions on Industrial Informaticss, arXiv:1909.06977, 2019.

[22] R. Monaco, C. Bergaentzle, P. S. Nielsen, and K. Sundsgaard, "Framing barriers for distribution grid digitalisation: A conceptual framework for policy recommendation," IEEE PES Innovative Smart Grid Technologies-Asia (ISGT Asia), pp. 1–5, 2023.

[23] H. Sæle, I. B. Sperstad, K. W. Hoiem, and V. Mathiesen, "Understanding barriers to utilising flexibility in operation and planning of the electricity distribution system–Classification frameworks with applications to Norway," Energy Policy, Elsevier, vol. 180, p. 113618, 2023.

[24] L. Feng et al., "Anomaly detection for electricity consumption in cloud computing: framework, methods, applications, and challenges," EURASIP Journal on Wireless Communications and Networking, Springer, no. 1, pp. 1–12, 2020.

[25] F. J. Sissine, "Energy Independence and Security Act of 2007: a summary of major provisions," Congressional Research Service Washington, DC, 2007.

[26] S. Brahma, R. Kavasseri, H. Cao, N. R. Chaudhuri, T. Alexopoulos, and Y. Cui, "Real-time identification of dynamic events in power systems using PMU data, and potential applications—models, promises, and challenges," IEEE transactions on Power Delivery, vol. 32, no. 1, pp. 294–301, 2016.

[27] G. Prettico, A. Marinopoulos, and S. Vitiello, "Guiding electricity distribution system investments to improve service quality: A European study," Utilities Policy, Elsevier, vol. 77, p. 101381, 2022.

[28] M. Shahraeini, M. H. Javidi, and Z. Haq, "Wide area measurement systems," Advanced topics in measurements, InTech Rijeka, Croatia, pp. 303–322, 2012.

[29] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," IEEE communications surveys & tutorials, IEEE, vol. 15, no. 1, pp. 5–20, 2012.

[30] H. Zareipour, "Digital power grid lessons from history to imagine the future," IEEE Power and Energy Magazine, IEEE, 2022.

[31] U. Häger, T. Wagner, C. Kittl, J. Jakob, and J. Hiry, "Digital Twins in Power Systems: A Proposal for a Definition," IEEE Power and Energy Magazine, vol. 22, no. 1, pp. 16–23, 2024.

[32] N. Cohn, S. B. Biddle, R. G. Lex, E. H. Preston, C. W. Ross, and Whitten, "On-line computer applications in the electric power industry," Proceedings of the IEEE, vol. 58, no. 1, pp. 78–87, 1970.

[33] B. Danette Allen, "Digital twins and living models at NASA," Digital Twin Summit, American Society of Mechanical Engineers, 2021.

[34] D. Gelernter, Mirror worlds: Or the day software puts the universe in a shoebox. How it will happen and what it will mean: Oxford University Press, 1993.

[35] M. W. Grieves, "Virtually indistinguishable: Systems engineering and PLM," IFIP International Conference on Product Lifecycle Management, Springer, 2012.

[36] E. Negri, L. Fumagalli, and M. Macchi, "A review of the roles of digital twin in CPS-based production systems," Procedia manufacturing, Elsevier, vol. 11, pp. 939–948, 2017.

[37] B. Rodič, "Industry 4.0 and the new simulation modelling paradigm," Organizacija, De Gruyter Poland, vol. 50, no. 3, pp. 193–207, 2017.

[38] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and US Air Force vehicles," 53rd AIAA/ASME/ASCE/AHS/ASC Structures, 1818, 2012.

[39] R. Rosen, G. von Wichert, G. Lo, and K. D. Bettenhausen, "About the importance of autonomy and digital twins for the future of manufacturing," Ifac-Papersonline, Elsevier, vol. 48, no. 3, pp. 567–572, 2015.

[40] S. Boschert and R. Rosen, "Digital twin—the simulation aspect," Mechatronic futures: Challenges and solutions for mechatronic systems and their designers, Springer, pp. 59–74, 2016.

[41] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," Transdisciplinary perspectives on complex systems: New findings and approaches, Springer, pp. 85–113, 2017.

[42] Z. Liu, N. Meyendorf, and N. Mrad, "The role of data fusion in predictive maintenance using digital twin," Handbook of Nondestructive Evaluation 4.0, Springer, no. 1, p. 20023, 2025.

[43] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," IEEE Access, vol. 8, pp. 108952–108971, 2020.

[44] Z. Bing et al., "Digital twin on concepts, enabling technologies, and applications," Journal of the Brazilian Society of Mechanical Sciences and Engineering, Springer, vol. 46, no. 7, p. 420, 2024.

[45] G. Schrotter and C. Hürzeler, "The digital twin of the city of Zurich for urban planning," PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science, vol. 88, no. 1, pp. 99–112, 2020.

[46] Y. Xu, Y. Sun, X. Liu, and Y. Zheng, "A digital-twin-assisted fault diagnosis using deep transfer learning," IEEE Access, vol. 7, pp. 19990–19999, 2019.

[47] M. Joordens and M. Jamshidi, "On the development of robot fish swarms in virtual reality with digital twins," 13th Annual Conference on System of Systems Engineering (SoSE), IEEE, pp. 411–416, 2018.

[48] V. J. Mawson and B. R. Hughes, "The development of modelling tools to improve energy efficiency in manufacturing processes and systems," Journal of Manufacturing Systems, Elsevier, vol. 51, pp. 95–105, 2019.

[49] T. Plank, "Digital Twins: The 4 types and their characteristics," Tributech, 2019. [Online]. Available: https://www.tributech.io/blog/the-4-types-of-digital-twins

[50] C. Jackson, "What is a Digital Twin?," Lifecycle Insights, 2018. https:// www.lifecycleinsights.com/tech-guide/digital-twins/

[51] A. Sharma, E. Kosasih, J. Zhang, A. Brintrup, and A. Calinescu, "Digital twins: State of the art theory and practice, challenges, and open research questions," Journal of Industrial Information Integration, Elsevier, vol. 30, p. 100383, 2022.

[52] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital Twin in manufacturing: A categorical literature review and classification," Ifac-Papersonline, Elsevier, vol. 51, no. 11, pp. 1016–1022, 2018.

[53] B. Tekinerdogan and C. Verdouw, "Systems architecture design pattern catalog for developing digital twins," Sensors, MDPI, vol. 20, no. 18, p. 5103, 2020.

[54] Häger, Ulf, et al., "The Digital Twin in the Network and Electricity Industry," VDE ETG, May. 2023.

[55] U. Häger, T. Wagner, C. Kittl, J. Jakob, and J. Hiry, "Digital Twins in Power Systems – A proposal for a definition," IEEE Power and Energy Magazine, IEEE, vol. 22, 2024.

[56] I. Kamwa and B. Badrzadeh, "From Buzzword To Solutions: Digital Twins in Power Systems," IEEE Power and Energy Magazine, IEEE, vol. 22, no. 1, pp. 4–11, 2024.

[57] A.Kummerow, S. Nicolai, C. Brosinsky, D. Westermann, A.Naumann, and M. Richter, "Digital-Twin based Services for advanced Monitoring and Control of future power systems," IEEE Power & Energy Society General Meeting (PESGM), 2020.

[58] C. Brosinsky, D. Westermann, and R. Krebs, "Recent and prospective developments in power system control centers: Adapting the digital twin technology for application in power system control centers," IEEE International Energy Conference (ENERGYCON), 2018.

[59] H. Mohammadi Moghadam, H. Foroozan, M. Gheisarnejad, and M.-H. Khooban, "A survey on new trends of digital twin technology for power systems," Journal of Intelligent & Fuzzy Systems, IEEE, vol. 41, no. 2, pp. 3873–3893, 2021.

[60] A. Baldassarre, A. Ceruti, D. N. Valyou, and P. Marzocca, "Towards a digital twin realization of the blade system design study wind turbine blade," Wind and Structures, Techno-Press, vol. 28, no. 5, pp. 271–284, 2019.

[61] P. Jain, J. Poon, J. P. Singh, C. Spanos, S. R. Sanders, and S. K. Panda, "A digital twin approach for fault diagnosis in distributed photovoltaic systems," IEEE Transactions on Power Electronics, vol. 35, no. 1, pp. 940–956, 2019.

[62] E. Espejo, F. Segundo Sevilla, P. Korba, Ed., Monitoring and Control of Electrical Power Systems Using Machine Learning Techniques: Elsevier, 2023.

[63] Y. Zhong, Wei Zhang, X. Ha, Q. Chen, J. Huang, and K. Yan, "Innovative digital twin platform construction for smart grid system," IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), 2021.

[64] M. Borth, J. Verriet, and G. Muller, "Digital twin strategies for SoS 4 challenges and 4 architecture setups for digital twins of SoS," 14th annual conference system of systems engineering (SoSE), IEEE, 2019.

[65] W. Danilczyk, Y. Lindsay Sun, and H. He, "Smart grid anomaly detection using a deep learning digital twin," 52nd North American Power Symposium (NAPS), IEEE, 2021.

[66] J. Akram, A. Tahir, H. S. Munawar, A. Akram, A. Z. Kouzani, and M. P. Mahmud, "Cloud- and fog-integrated smart grid model for efficient resource utilisation," Sensors, MDPI, vol. 21, no. 23, p. 7846, 2021.

[67] Y. Lin, C. Cheng, F. Xiao, K. Alsubhi, and H. M. A. Aljahdali, "A DAG-based cloud-fog layer architecture for distributed energy management in smart power grids in the presence of PHEVs," Sustainable Cities and Society, Elsevier, vol. 75, p. 103335, 2021.

[68] C. Köhler, R. Kersten, and M. Schöpf, "Cloud-Based Digital Twin for Distribution Grids: What Is Already Available Today," IEEE Power and Energy Magazine, IEEE, vol. 22, no. 1, pp. 72–80, 2024.

[69] C. Brosinsky, D. Westermann, and R. Krebs, "Recent and prospective developments in power system control centers: Adapting the digital twin technology for application in power system control centers," international energy conference (ENERGYCON), IEEE, pp. 1–6, 2018.

[70] C. Rehtanz, U. Häger, and C.-C. Liu, "Digital Twin: From Buzzword To Solutions," Power and Energy Magazine, IEEE, vol. 22, no. 1, pp. 14–15, 2024.

[71] P. Palensky, P. Mancarella, T. Hardy, and M. Cvetkovic, "Cosimulating Integrated Energy Systems With Heterogeneous Digital Twins: Matching a Connected World," IEEE Power and Energy Magazine, vol. 22, no. 1, pp. 52–60, 2024.

[72] S. Singh et al., "Data management for developing digital twin ontology model," Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, vol. 235, no. 14, pp. 2323–2337, 2021.

[73] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," Computer science review, Elsevier, vol. 17, pp. 70–81, 2015.

[74] D. Klein, P. Tran-Gia, and M. Hartmann, "Big data," Informatik-Spektrum, Springer, vol. 36, pp. 319–323, 2013.

[75] B. P. Bhattarai et al., "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," IET Smart Grid, Wiley Online Library, vol. 2, no. 2, pp. 141–154, 2019.

[76] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," Renewable and sustainable energy reviews, Elsevier journal, vol. 56, pp. 215–225, 2016.

[77] J. Tang and H. Sui, "Application technology of big data in smart grid and its development prospect," International Conference on Computer Technology, Electronics and Communication (ICCTEC), IEEE, 2017.

[78] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," Energy informatics, Springer, vol. 1, no. 1, pp. 1–24, 2018.

[79] H.-N. Dai, R. C.-W. Wong, H. Wang, Z. Zheng, and A. V. Vasilakos, "Big data analytics for large-scale wireless networks: Challenges and opportunities," ACM computing surveys (CSUR), ACM New York, NY, USA, vol. 52, no. 5, pp. 1–36, 2019.

[80] B. Caesar, A. Hänel, E. Wenkler, C. Corinth, S. Ihlenfeldt, and A. Fay, "Information model of a digital process twin for machining processes," 25th IEEE international conference on emerging technologies and factory automation (ETFA), 2020.

[81] A. E. Saldaña-González, A. Sumper, M. Aragüés-Peñalba, and M. Smolnikar, "Advanced distribution measurement technologies and data applications for smart grids: A review," Energies, MDPI, vol. 13, no. 14, p. 3730, 2020.

[82] Siemens AG, Siemens Electrical Digital Twin, 2018. [Online]. Available: siemens.com/electrical-digital-twin

[83] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the Digital Twin: A systematic literature review," CIRP Journal of Manufacturing Science and Technology, Elsevier, vol. 29, pp. 36–52, 2020.

[84] A. M. Madni, C. C. Madni, and S. D. Lucero, "Leveraging digital twin technology in model-based systems engineering," Systems, MDPI, vol. 7, no. 1, p. 7, 2019.

[85] H. A. Park, G. Byeon, W. Son, H. C. Jo, J. Kim, and S. Kim, "Digital Twin for Operation of Microgrid: Optimal Scheduling in Virtual Space of Digital Twin. Energies 2020, 13, 5504," Energies, MDPI, vol. 13, 2020.

[86] S. V. Nath, P. van Schalkwyk, and D. Isaacs, "Building industrial digital twins: Design, develop, and deploy digital twin solutions for real-world industries using Azure digital twins," Packt Publishing Ltd, 2021.

[87] M. Endsley, "Design and evaluation for situation awareness enhancement," Proceedings of the Human Factors Society annual meeting, Sage Publications Sage CA: Los Angeles, CA, vol. 2, 1988.

[88] R. Parasurman, T. Bahari, J. Deaton, J. Morrison, and M. Barnes, "Theory and design of adaptive automation in aviation systems," The Catholic University of America, Washington, DC, 1992.

[89] T. S. Carretta, D. C. Perry Jr, and M. J. Ree, "Prediction of situational awareness in F-15 pilots," The International Journal of Aviation Psychology, Taylor & Francis, vol. 6, no. 1, pp. 21–41, 1996.

[90] C. Zimmerman, "Ten strategies of a world-class cybersecurity operations centre. the mitre corporation," 2014.

[91] A. Muir and J. Lopatto, "Final report on the August 14, 2003 blackout in the United States and Canada: causes and recommendations," U.S. Department of Energy Office of Scientific and Technical Information, 2004.

[92] M. Endsley E. Connors, "Enhancing situation awareness in power system control centers," IEEE International Multi-Disciplinary Conference on Cognitive Methods, 2013.

[93] P. Hines, J. Apt, and S. Talukdar, "Large blackouts in North America: Historical trends and policy implications," Energy Policy, Elsevier, vol. 37, no. 12, pp. 5249–5259, 2009.

[94] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," Human factors, SAGE Publications Sage CA: Los Angeles, CA, vol. 37, no. 1, pp. 32–64, 1995.

[95] Z. Dong, T. Xu, Y. Li, P. Feng, X. Gao, and X. Zhang, "Review and application of situation awareness key technologies for smart grid," IEEE Conference on Energy Internet and Energy System Integration (EI2), 2017.

[96] A. Shahsavari, M. Farajollahi, E. Stewart, E. Cortez, H. Mohsenian-Rad†, "A machine learning approach to event analysis in distribution feeders using distribution synchrophasors," International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA), IEEE, 2019.

[97] A. von Meier, D. Culler, A. McEachern, and R. Arghandeh, "Micro-synchrophasors for distribution systems," IEEE, 2014.

[98] M. Babakmehr, M. G. Simões, M. B. Wakin, and F. Harirchi, "Compressive sensing-based topology identification for smart grids," IEEE Transactions on Industrial Informatics, vol. 12, no. 2, pp. 532–543, 2016.

[99] J. Wang, "Power system short-term load forecasting," 5th international conference on machinery, materials and computing technology (ICMMCT 2017), Atlantis Press, 2017.

[100] X. Xu, "Short-term load forecasting of power system," Advances in Materials, Machinery, Electrical Engineering (AMMEE), 2017.

[101] A. D. Lotufo and C. R. Minussi, "Electric power systems load forecasting: A survey," PowerTech Budapest 99. Abstract Records.(Cat. No. 99EX376), IEEE, p. 36, 1999.

[102] S. Liao, Y. Liu, J. Xu, L. Jia, D. Ke, and X. Jiang, "Data-Driven Real-Time Congestion Forecasting and Relief With High Renewable Energy Penetration," IEEE Transactions on Industrial Informatics, 2024.

[103] K. Zor, O. Timur, and A. Teke, "A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting," 6th international youth conference on energy (IYCE), IEEE, 2017.

[104] J. Walther, D. Spanier, N. Panten, and E. Abele, "Very short-term load forecasting on factory level–A machine learning approach," Procedia CIRP, Elsevier, vol. 80, pp. 705–710, 2019.

[105] L. Botman, J. Lago, T. Becker, O Agudelo, K. Vanthournout, and B. De Moo, "A scalable method for probabilistic short-term forecasting of individual households consumption in low voltage grids," IEEE PES Grid Edge Technologies Conference & Exposition (Grid Edge), 2023.

[106] P. Qingle and Z. Min, "Very short-term load forecasting based on neural network and rough set," international conference on intelligent computation, IEEE, 2010.

[107] Nadtoka and M. Balasim, "Mathematical modelling and short-term forecasting of electricity consumption of the power system, with due account of air temperature and natural illumination, based on support vector machine and particle swarm," Procedia engineering, vol. 129, pp. 657–663, 2015.

[108] C. J. Bennett, R. A. Stewart, and J. W. Lu, "Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system," Energy, Elsevier, vol. 67, pp. 200–212, 2014.

[109] M. Reis, A. Garcia, R. Bessa, "A scalable load forecasting system for low voltage grids," IEEE Manchester PowerTech, 2017.

[110] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," IEEE Transactions on power systems, vol. 16, no. 1, pp. 44–55, 2001.

[111] Z. Cao, C. Wan, Z. Zhang, F. Li, and Y. Song, "Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting," IEEE Transactions on power systems, vol. 35, no. 3, pp. 1881–1897, 2019.

[112] D. Samariya and A. Thakkar, "A comprehensive survey of anomaly detection algorithms," Annals of Data Science, Springer, vol. 10, no. 3, pp. 829–850, 2023.

[113] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: A survey," ACM Computing Surveys, vol. 14, p. 15, 2007.

[114] Q. Wen, J. Gao, X. Song, L. Sun, H. Xu, S. Zhu, "Robust STL: A robust seasonal-trend decomposition algorithm for long time series," Proceedings of the AAAI conference on artificial intelligence, vol. 01, 2019.

[115] M. Hojabri, S. Nowak, and A. Papaemmanouil, "ML-based intermittent fault detection, classification, and branch identification in a distribution network," Energies, MDPI, vol. 16, no. 16, p. 6023, 2023.

[116] S. J. Pinto, P. Siano, and M. Parente, "Review of cybersecurity analysis in smart distribution systems and future directions for using unsupervised learning methods for cyber detection," Energies, vol. 16, no. 4, p. 1651, 2023.

[117] A. Althobaiti, A. Jindal, A. K. Marnerides, and U. Roedig, "Energy theft in smart grids: a survey on data-driven attack strategies and detection methods," IEEE Access, vol. 9, pp. 159291–159312, 2021.

[118] I. Semertzis, H. Goyel, V. Rajkumar, A. Presekal, A. Ştefanov, and P. Palensky, "Towards Real-Time Distinction of Power System Faults and Cyber Attacks," IEEE Power & Energy Society General Meeting (PESGM), 2023.

[119] I. Natgunanathan, V. Mak-Hau, S. Rajasegarar, and A. Anwar, "Deakin microgrid digital twin and analysis of AI models for power generation prediction," Energy Conversion and Management: Elsevier, vol. 18, p. 100370, 2023.

[120] R. Fontugne et al., "Strip, bind, and search: a method for identifying abnormal energy consumption in buildings," Association for Computing Machinery (ACM), IEEE, pp. 129–140, 2013.

[121] Z. Wang, Q. Cui, Z. Gong, L. Shi, J. Gao, and J. Zhong, "Anomaly Identification for Photovoltaic Power Stations Using a Dual Classification System and Gramian Angular Field Visualization," Processes, MPDI, vol. 12, no. 4, p. 690, 2024.

[122] H.-S. Nam, Y.-K. Jeong, and J. W. Park, "An anomaly detection scheme based on LSTM autoencoder for energy management," international conference on information and communication technology convergence (ICTC), IEEE, pp. 1445–1447, 2020.

[123] S. Lee et al., "Smart metering system capable of anomaly detection by bi-directional LSTM autoencoder," IEEE International Conference, pp. 1–6, 2020.

[124] L. Yang, Y. Zhai, and Z. Li, "Deep learning for online AC false data injection attack detection in smart grids: An approach using LSTM-autoencoder," Journal of Network and Computer Applications, vol. 193, p. 103178, 2021.

[125] Y. Ma, D. Oslebo, A. Maqsood, and K. Corzine, "DC fault detection and pulsed load monitoring using wavelet transform-fed LSTM autoencoders," IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 9, no. 6, pp. 7078–7087, 2020.

[126] A. Mahi-Al-rashid, F. Hossain, A. Anwar, and S. Azam, "False data injection attack detection in smart grid using energy consumption forecasting," Energies, MDPI, vol. 15, no. 13, p. 4877, 2022.

[127] P.K. Lim and D.S. Dorr, "Understanding and resolving voltage sag related problems for sensitive industrial customers," IEEE Power Engineering Society Winter Meeting. Conference Proceedings, 2000.

[128] S. Mishra, C. N. Bhende, and B. K. Panigrahi, "Detection and classification of power quality disturbances using S-transform and probabilistic neural network," IEEE transactions on Power Delivery, vol. 23, no. 1, pp. 280–287, 2007.

[129] T. Radil, P. M. Ramos, F. M. Janeiro, and A. C. Serra, "PQ monitoring system for real-time detection and classification of disturbances in a single-phase power system," IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 8, pp. 1725–1733, 2008.

[130] A. Anwar and A. N. Mahmood, "Anomaly detection in electric network database of smart grid: Graph matching approach," Electric Power Systems Research, Elsevier, vol. 133, pp. 51–62, 2016.

[131] M. Zhou and P. Musilek, "Real-time anomaly detection in distribution grids using long short term memory network," in 2021 IEEE Electrical Power and Energy Conference (EPEC), pp. 208–213.

[132] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," IEEE transactions on knowledge and data engineering, vol. 24, no. 5, pp. 823–839, 2010.

[133] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," National science review, IEEE, vol. 1, no. 2, pp. 293–314, 2014.

[134] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," Journal of the Royal Statistical Society Series B: Statistical Methodology, Oxford University Press (OUP), UK, vol. 70, no. 5, pp. 849–911, 2008.

[135] F. Sythoff, Why Smart Meter Programs Run into Roadblocks without Real-Time Data. [Online]. Available: https://www.greenbird.com/resources/why-smart-meter-programs-run-into-roadblocks-without-real-time-data

[136] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, vol. 1, no. 1, pp. 1–22, 2016.

[137] E.-A.-U. An, "Smart meters and smart meter systems: A metering industry perspective," Washington, DC, USA, Edison Elect. Inst., White Paper, EEI, 2011.

[138] R. Balouchi, M. Weissenstein, W. Wellßow, "Pseudo-worst-case forecast for a preventive control in LV smart grids," NEIS 2020; Conference on Sustainable Energy Supply and Energy Storage Systems, IEEE, 2020.

[139] R. Balouchi, U. Häger, W. Wellßow, "Pseudo-Worst-Case Forecast with Neural Networks in Low Voltage Grids," IEEE PowerTech, 2023.

[140]DIN EN 50160:2020-11, Merkmale der Spannung in öffentlichen Elektrizitätsversorgungsnetzen; Deutsche Fassung EN_50160:2010_+ Cor.:2010_+ A1:2015_+ A2:2019_+ A3:2019, Berlin.

[141] M. Weisenstein, "Beitrag zur Netzautomatisierung auf Niederspannungsebene unter Berücksichtigung spärlicher Datenlage und-übertragungsrate," Dissertation, Technical university of Kaiserslautern, 2022.

[142] R. Balouchi, R. Palaniappan,U. Häger, C. Rehtanz, "Data-driven Approaches for Anomaly Detection in Low-Voltage Grid Net Power," IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2024.

[143] R. Balouchi, R. Palaniappan, U. Häger, C. Rehtanz, "Anomaly Detection in Low-Voltage Grids with LSTM Autoencoders: A Study on Future Scenario Impacts," IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2024.

[144] Isolation forest: IEEE, 2008.

[145] Streaming linear regression on spark MLlib and MOA, 2015.

[146] A. D. Masood, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine learning supervised algorithms of gene selection: A review," Machine Learning, vol. 62, no. 03, 2020.

[147] The research of regression model in machine learning field: EDP Sciences, 2018.

[148] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure," Technometrics, vol. 25, no. 2, pp. 165–172, 1983.

[149] Y. Zhou, J. Zhao, Y. Song, J. Sun, H. Fu, and M. Chu, "A seasonal–trend-decomposition-based voltage-source-inverter open-circuit fault diagnosis method," IEEE Transactions on Power Electronics, vol. 37, no. 12, pp. 15517–15527, 2022.

[150] A. Mishra, R. Sriharsha, and S. Zhong, "Online STL: Scaling time series decomposition by 100x," arXiv preprint arXiv:2107.09110, 2021.

[151] P. Mobtahej, X. Zhang, M. Hamidi, and J. Zhang, "An LSTM-Autoencoder Architecture for Anomaly Detection Applied on Compressors Audio Data," Computational and Mathematical Methods, Wiley-Blackwell Publishing Ltd (United Kingdom), no. 1, p. 3622426, 2022.

[152] Shaun Turney, Central Limit Theorem | Formula, Definition & Examples. [Online]. Available: https://www.scribbr.com/statistics/central-limit-theorem/

[153] J. W. Tukey, "Exploratory data analysis," Reading/Addison-Wesley, 1977.

[154] Deutscher Wetterdienst. [Online]. Available: https://www.dwd.de/

[155] D. Kim, A. Turlikov, G. Georgiev, and T. Dedova, "LoRaWAN Optimization for Voltage Monitoring," IEEE Access, 2024.

[156] S. Meinecke et al., "Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis," Energies, MPDI, vol. 13, no. 12, p. 3290, 2020.

[157] D. Sarajlić and C. Rehtanz, "Low voltage benchmark distribution network models based on publicly available data," IEEE PES Innovative Smart Grid Technologies Conference Europe, 2019.

[158] S. Meinecke et al., "SimBench Documentation-Documentation Version EN-1.0. 0," Technical Report, 2020.

[159] Siemens AG, Intelligente Ortsnetzstationen für zukunftssichere Energieverteilung: Das modulare 3-Stufen-Konzept von Siemens. [Online]. Available: https://assets.new.siemens.com/siemens/assets/api/uuid:0f81d76b2b388548add30d15420dc38277ba6480/version:1682527323/a-intelligent-transformer-substation-de.pdf

[160] W. S. McCulloch and W. Pitts, “A Logical Calculus of the Ideas Immanent in Nervous Activity,” Bulletin of Mathematical Biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[161] W. H Woodall, and, B. M. Adams; "The Statistical Design of CUSUM Charts", Quality Engineering, 5(4), 559-570, 1993.

[162] Sarem, “Probabilistic CUSUM for Change Point Detection,” Aug. 4, 2022. Available: https://sarem-seitz.com/posts/probabilistic-cusum-for-change-point detection.html

[163] E. Parzen, “On estimation of a probability density function and mode,” The Annals of Mathematical Statistics, vol. 33, no. 3, pp. 1065–1076, 1962.

[164] S. Kullback and R. A. Leibler, “On information and sufficiency,” The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.