FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Enhancing Data Availability in Distribution Networks through Synthetic Historical Data Reconstruction

Rui Miguel Mendes Brito

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Ricardo Bessa

October 12, 2024

Resumo

Desde a sua criação, as redes de baixa tensão têm-se caracterizado por uma notável falta de precisão e fiabilidade na provisão dos seus parâmetros definidores. Exemplo disso é a inexistência de uma caraterização elétrica fidedigna dos seus componentes e topologia, e a ausência de telemetria. Contudo, a sua procura crescente, impulsionada pelo aumento de carregadores de veículos elétricos (VE), exige uma transformação substancial da arquitetura da rede de baixa tensão (BT). Simultaneamente, a proliferação significativa de fontes de microgeração e geração distribuída intensifica uma coordenação meticulosa por parte do operador do sistema de distribuição (*Distribution System Operator* - DSO) para assim se otimizar a gestão dos recursos.

Desta forma, as técnicas convencionais de análise de rede não podem ser consideradas perante a ausência de representações precisas do seu modelo físico. Por outro lado, no advento da massificação da utilização de contadores inteligentes (*smart meters*), os enormes volumes de dados gerados apresentam-se, muitas vezes, como a única alternativa a considerar para o desenvolvimento de ferramentas de monitorização e gestão da rede BT, naturalmente explorados por algoritmos de aprendizagem automática.

Com o crescente reconhecimento da importância desses dados, surgem modelos de negócios orientados à sua monetização. Neste contexto, emerge um trade-off técnico-económico significativo entre a aquisição de dados reais e a sua reconstrução sob a denominação de dados sintéticos. Ademais, a obtenção de dados reais pode enfrentar desafios relacionados com a privacidade dos dados dos medidores inteligentes, a disponibilidade limitada e a potencial falta de granularidade. Em contrapartida, a geração de dados sintéticos oferece uma alternativa inovadora, permitindo superar tais obstáculos, desde que salvaguardadas as devidas condições quanto à sua representatividade.

O desafio proposto neste trabalho envolve o desenvolvimento de estratégias para a reconstrução de dados ausentes utilizando diferentes técnicas de aprendizagem automática, começando com um conjunto inicial de dados limitados. Para tal, destacam-se três principais objetivos de caráter geral: i) garantir a similaridade entre as séries temporais geradas e as reais, ii) assegurar a preservação das propriedades das distribuições de probabilidade, e iii) manter a coerência com o modelo físico que os dados reconstruídos visam representar. Neste estudo, também serão exploradas as implicações da integração de geração distribuída e cargas de alta potência na rede de baixa tensão, bem como os seus efeitos nos modelos de reconstrução.

Palavras-chave: Baixa tensão, rede inteligente, contadores inteligentes, geração de dados sintéticos, aprendizagem automática, reconstrução de histórico

Abstract

Since their inception, low-voltage networks have been distinguished by a conspicuous deficiency in precision and reliability with regard to the specification of their defining parameters. One example of this is the lack of reliable electrical characterisation of their components and topology, as well as the absence of telemetry. However, the increasing demand for low-voltage networks, driven by the expansion of electric vehicle (EV) chargers, necessitates a significant transformation of the existing network architecture. Concurrently, the considerable increase in microgeneration and distributed generation sources has led to a greater need for meticulous coordination by the Distribution System Operator (DSO) in order to optimise resource management.

It is, therefore, evident that conventional network analysis techniques cannot be considered in the absence of accurate representations of the physical model. Conversely, the advent of widespread smart meter usage has led to the generation of vast quantities of data, which represents a potential avenue for the development of LV network monitoring and management tools. This data can be harnessed by machine learning algorithms, which are becoming increasingly prevalent in this field.

As the importance of this data is increasingly recognised, business models are emerging with the objective of monetising it. In this context, there is a significant technical and economic trade-off between acquiring data in its original form and reconstructing it as synthetic data. Moreover, the acquisition of genuine data may encounter obstacles pertaining to the confidentiality of smart meter data, restricted accessibility, and a potential deficiency in granularity. Conversely, the generation of synthetic data provides an innovative alternative, enabling these obstacles to be overcome, provided the appropriate conditions are met regarding its representativeness.

The challenge proposed in this work is to develop strategies for reconstructing missing data using different machine-learning techniques, beginning with an initial set of limited data. In order to achieve these goals, three main general objectives have been identified: firstly, to ensure similarity between the generated time series and the real ones; secondly, to ensure that the properties of probability distributions are preserved; and thirdly, to maintain consistency with the physical model that the reconstructed data is intended to represent. In addition, this study will also explore the implications of integrating distributed generation and high-power loads into the low-voltage network, as well as their effects on the reconstruction models.

Keywords: Low voltage, smart grid, smart meters, synthetic data generation, machine learning, historical reconstruction

Sustainable Development Goals (SDGs) of the United Nations

This work addresses critical aspects of modern energy systems, focusing on enhancing the management of low-voltage networks to support sustainable development. Utilizing advanced techniques such as synthetic data generation and machine learning contributes to several sustainable development goals (SDGs).

SDG	Target	Contribution	Performance Indicator and Metrics
7	7.1	Enhances management of low-voltage networks to support integration of renewable energy sources and electric vehicles (EVs).	Improved reliability and efficiency metrics of low-voltage networks, increased penetration of renewable energy and EVs.
7	7.2	Promotes clean energy usage through better network management and data-driven decision-making.	Percentage increase in renewable energy share within the energy mix.
7	7.3	Improves energy efficiency with synthetic data generation and predictive modelling.	Rate of improvement in energy efficiency, reduction in energy losses.
9	9.1	Contributes to building resilient and sustainable electric distribution networks.	Network reliability indices (SAIDI, SAIFI), infrastructure resilience measures.
9	9.4	Optimizes network performance and reliability through data reconstruction and predictive modelling.	Metrics on infrastructure upgrades, operational cost reduction, and resource-use efficiency improvement.
11	11.6	Enhances efficiency and reliability of urban low-voltage networks.	Reduction in urban carbon footprint, improved air quality indices.
11	11.B	Supports better planning and integration of sustainable energy solutions.	Number of cities adopting sustainable energy plans, policy implementation metrics.
12	12.2	Promotes sustainable management of energy resources.	Improvement in resource-use efficiency, reduction in energy wastage.
12	12.6	Enables accurate data-driven decision-making for sustainable practices.	Number of companies adopting sustainable energy practices, compliance rates.
13	13.1	Enhances resilience of energy infrastructure to climate-related disruptions.	Metrics on infrastructure resilience, reduction in climate-related energy disruptions.
13	13.3	Supports adaptation to increased renewable integration and changing demand patterns.	Improvement in institutional capacity, number of educational programs and training conducted.

Table 1: Contributions to the Sustainable Development Goals (SDGs)

SDG 7: Affordable and Clean Energy

This research enhances the management of low-voltage networks, essential for integrating renewable energy sources and electric vehicles (EVs). By developing strategies for synthetic data generation, the research supports these networks' reliable and efficient operation, promoting clean energy usage and improving energy efficiency.

SDG 9: Industry, Innovation, and Infrastructure

The work on reconstructing historical data using machine learning techniques directly contributes to building a more resilient and sustainable infrastructure for electric distribution networks. Innovations in data reconstruction and predictive modelling optimize network performance and reliability, which are essential for modern infrastructure.

SDG 11: Sustainable Cities and Communities

By improving the efficiency and reliability of low-voltage networks, particularly in urban areas where EV adoption and distributed generation are growing, this research aids in reducing the environmental impact of cities. Enhanced data availability and predictive modelling support better planning and integration of sustainable energy solutions in urban settings.

SDG 12: Responsible Consumption and Production

This thesis promotes the sustainable management of energy resources by enabling more accurate data-driven decision-making in electrical distribution. Synthetic data generation improves resource management and operational efficiency, leading to more responsible consumption and production patterns.

SDG 13: Climate Action

This work supports climate action by ensuring that energy networks can adapt to and mitigate the impacts of increased renewable integration and changing demand patterns.

Agradecimentos

Gostaria de expressar a minha gratidão a todas e todos que de alguma forma contribuíram, não só para o desenvolvimento deste trabalho, mas também para o meu desenvolvimento académico, pessoal e profissional ao longo destes 5 anos.

Em primeiro lugar, gostaria de agradecer aos meus pais, ao meu irmão e a toda a minha família por todo o apoio que me deram e pela possibilidade de poder cumprir mais esta etapa da minha vida sem que me faltasse nada. Não há palavras que cheguem para mostrar a minha gratidão.

Gostaria também de agradecer ao meu orientador Ricardo Bessa pela oportunidade que me proporcionou de trabalhar com uma equipa multicultural no INESC TEC. Gostaria também de expressar a minha gratidão ao Gil Sampaio por todo o apoio dado ao longo destes meses, desde a primeira fase da dissertação até ao último dia com as revisões sucessivas do meu trabalho.

À Sandra, por toda a paciência em lidar com os meus desabafos e por todo o apoio que me deu ao longo deste ano. Obrigado por fazeres parte da minha vida.

Aos meus camaradas de trabalho e amigos Tiago, Sérgio, Nuno, Fábio e Renato, pelo companheirismo, as jantaradas e os bons momentos que foram fundamentais para tornar toda esta jornada de 5 anos uma experiência inesquecível que vai deixar saudades.

Rui Miguel Mendes Brito

"The only way to discover the limits of the possible
is to go beyond them into the impossible."

Arthur Charles Clarke

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Objectives . . . 2
- 1.3 Document Structure . . . 3

2 Background and State of the Art **4**
- 2.1 Smart Grids . . . 4
  - 2.1.1 Smart Meters . . . 5
  - 2.1.2 Consumer's Smart Meters use-cases . . . 6
- 2.2 Smart Meters Synthetic Data Generation . . . 8
  - 2.2.1 Stochastic Approaches . . . 9
  - 2.2.2 Machine learning Approaches . . . 10
- 2.3 Theoretical Background . . . 13
  - 2.3.1 Artificial Neural Networks . . . 13
  - 2.3.2 Recurrent Neural Networks . . . 16
  - 2.3.3 Long Short-Term Memory (LSTM) networks . . . 17

3 Methodology **19**
- 3.1 Data Reconstruction . . . 19
- 3.2 Implemented algorithms . . . 20
  - 3.2.1 Baseline time series Reconstruction . . . 20
  - 3.2.2 Recurrent Neural Network (RNN) . . . 20
  - 3.2.3 Incorporation of Equipment-Related Information . . . 21
  - 3.2.4 Statistical Preserved Power Reconstruction (SPPR) . . . 21
- 3.3 Features and targets . . . 22
- 3.4 Train, Validation and Test Split . . . 23

4 Results **25**
- 4.1 Data overview . . . 25
  - 4.1.1 Data Collection . . . 25
  - 4.1.2 Data Transformation . . . 26
- 4.2 Evaluation metrics . . . 27
- 4.3 Probability distribution properties . . . 28
- 4.4 Voltage Time Series Reconstruction . . . 30
  - 4.4.1 Single-meter Reconstruction . . . 30
  - 4.4.2 Single-phase Reconstruction . . . 35
  - 4.4.3 Full Grid Reconstruction . . . 37
- 4.5 Power Injection Time series Reconstruction . . . 39

4.5.1 Single-meter Reconstruction 39
4.5.2 Single-Phase Reconstruction 43
4.5.3 Full Grid Reconstruction 45
4.6 Reconstruction with equipment knowledge 46
4.7 Voltage reconstruction considering statistical properties 51
4.8 Power reconstruction considering statistical properties 60
4.9 Physical properties Analysis 73

5 Conclusions and Future Work **76**
5.1 Conclusions 76
5.2 Future Work 77

List of Figures

2.1 Advanced Metering Infrastructure [2] . . . 6
2.2 Operations executed in each neuron of the Neural Network [26] . . . 13
2.3 Comonly used activation functions . . . 15
2.4 Feedforward Neural Network . . . 16
2.5 Recurrent Neural Network [35] . . . 16
2.6 Block diagram of the LSTM recurrent network "cell." [37] . . . 18

3.1 Linear Regression Data Split . . . 24
3.2 Neural Network Data Split . . . 24

4.1 Traditional Portuguese Grid [1] . . . 26
4.2 Method 1 . . . 27
4.3 Method 3 . . . 27
4.4 Error for each bus using LR forVoltage reconstruction . . . 38
4.5 Error for each bus using LSTMs for Voltage reconstruction . . . 38
4.6 Error for each bus using LR for Active Power injection reconstruction . . . 46
4.7 Error for each bus using LSTMs for Active Power injection reconstruction . . . 46
4.8 Power Injection Baseline Reconstruction . . . 47
4.9 Power Injection Reconstruction with EV Charger Consumption Information . . . 47
4.10 PV Generation Reconstrunction . . . 48
4.11 Baseline Reconstruction bus c_29 . . . 49
4.12 Baseline Reconstruction with EV information bus c_29 . . . 49
4.13 Baseline Reconstruction with PV information bus c_29 . . . 50
4.14 Baseline Reconstruction with EV and PV information bus c_29 . . . 50
4.15 Voltage Reconstruction Bus a_7 . . . 51
4.16 Probability Density Function comparison Bus a_7 . . . 52
4.17 ACF and PACF comparison Bus a_7 . . . 53
4.18 Serie decomposition Bus a_7 . . . 55
4.19 Voltage Reconstruction Bus a_29 . . . 56
4.20 Probability Density Function comparison Bus a_29 . . . 57
4.21 ACF and PACF comparison Bus a_29 . . . 58
4.22 Serie decomposition Bus a_29 . . . 59
4.23 Power Injection Baseline Reconstruction Bus a_7 . . . 60
4.24 SPPR model Bus a_7 . . . 60
4.25 Baseline model Probability Density Function comparison Bus a_7 . . . 62
4.26 SPPR model Probability Density Function comparison Bus a_7 . . . 62
4.27 Baseline model ACF and PACF comparison Bus a_7 . . . 63
4.28 SPPR model ACF and PACF comparison Bus a_7 . . . 63

4.29 Baseline Reconstrunciton Serie decomposition Bus *a*_7 64
4.30 SPPR Serie decomposition Bus *a*_7 65
4.31 Power Injection Baseline Reconstruction Bus *a*_29 66
4.32 Power Injection SPPR Bus *a*_29 66
4.33 Baseline model Probability Density Function comparison Bus *a*_29 68
4.34 SPPR Probability Density Function comparison Bus *a*_29 68
4.35 Baseline model ACF and PACF comparison Bus *a*_29 69
4.36 SPPR model ACF and PACF comparison Bus *a*_29 70
4.37 Baseline series decomposition Bus *a*_29 70
4.38 SPPR series decomposition Bus *a*_29 72
4.39 Sensitivity coefficients analysis for bus *a*_7 74
4.40 Sensitivity coefficients analysis for bus *a*_29 74

List of Tables

1 Contributions to the Sustainable Development Goals (SDGs) iii

3.1 Features and their descriptions 23

4.1 Comparison of LR Models for Single-meter Voltage Reconstruction (Part 1) . . . 31
4.2 Comparison of LR Models for Single-meter Voltage Reconstruction (Part 2) . . . 32
4.3 Comparison of LSTM Models for Single-meter Voltage Reconstruction (Part 1) . 33
4.4 Comparison of LSTM Models for Single-meter Voltage Reconstruction (Part 2) . 34
4.5 Error metrics for voltage time series reconstruction (single-meter) 34
4.6 Details of Model 1 for Single-phase Voltage Reconstruction 35
4.7 Details of Model 2 for Single-phase Voltage Reconstruction 36
4.8 Comparison of LSTM Models for Single-phase Voltage Reconstruction 36
4.9 Error Metrics for voltage time series reconstruction (single-phase) 36
4.10 Comparison of LR Models for Full Grid Voltage Reconstruction 37
4.11 Comparison of LSTM Models for Full Grid Voltage Reconstruction 37
4.12 Error Metrics for voltage time series reconstruction (Full Grid) 38
4.13 Comparison of LR Models for Single-meter Active Power Reconstruction (Part 1) 39
4.14 Comparison of LR Models for Single-meter Active Power Reconstruction (Part 2) 40
4.15 Comparison of LSTM Models for Single-meter Active Power Reconstruction (Part 1) 41
4.16 Comparison of LSTM Models for Single-meter Active Power Reconstruction (Part 2) 42
4.17 Error Metrics for Power injection time series reconstruction (single-meter) 42
4.18 Comparison of LR Models for Single-phase Active Power Reconstruction 43
4.19 Comparison of LSTM Models for Single-phase Active Power Reconstruction . . 44
4.20 Error Metrics for Power Injection Time Series Reconstruction (single-phase) . . . 44
4.21 Comparison of LR Models for Full Grid Active Power Reconstruction 45
4.22 Comparison of LSTM Models for Full Grid Active Power Reconstruction 45
4.23 Error Metrics for voltage time series reconstruction (Full Grid) 46
4.24 Error Metrics for Power Injection Time Series Reconstruction with and without EV charging power 47
4.25 Error Metrics for Photovoltaic Generation Reconstruction 48
4.26 Error Metrics for Power Injection Time Series Reconstruction 49
4.27 Error Metrics for Power Injection Time Series Reconstruction 50
4.28 Comparison of statistical properties between real and generated values from different models. 61
4.29 Comparison of statistical properties between real and generated values from different models. 67
4.30 Sensitivity Coefficients for Real data time series (First Five buses) 73

4.31 Sensitivity Coefficients for Generated data time series (First Five buses) 73

Abbreviations and Symbols

EV	Electric vehicle
DSO	Distribution System Operator
LV	Low Voltage
PV	Photovoltaic
AMI	Advanced Metering Infrastructure
ML	Machine Learning
GAN	Generative Adversarial Networks
VAE	Variational Autoencoders
NN	Neural Networks
ARIMA	Autoregressive Integrated Moving Average
DSF	Demand-side flexibility
RES	Renewable Energy Sources
EES	Energy Enhancement System
EDP	Eletricidade de Portugal
AI	Artificial Intelligence
SS	Secondary Substation

Chapter 1

Introduction

1.1 Context

The integration of advanced computational learning techniques for smart meter's historical data reconstruction within LV networks, particularly those accommodating electric vehicle (EV) chargers and distributed generation sources, poses a critical challenge for network operators. The surge in demand, driven by the proliferation of EV chargers and the increasing adoption of microgeneration, necessitates a significant transformation in the LV network architecture. Simultaneously, the imperative for meticulous coordination by Distribution System Operators (DSOs) intensifies to optimize grid management in the face of these dynamic changes.

However, the scarcity of granular electrical network data presents formidable challenges for network operators, hampering their ability to make informed decisions. Compounding this issue is the proprietary nature of critical data, raising concerns regarding accessibility, transparency, and equitable dissemination. These factors are pivotal for ensuring optimal performance and reliability of the electrical grid. Smart meter data raises privacy concerns due to the detailed nature of the information collected.

Smart meter data often includes personally identifiable information, such as the daily routine and occupancy patterns, which leads to potential risks such as unauthorized surveillance, behavioural profiling, and location-based inferences. Sharing data with third parties, regulatory compliance challenges, and the need for effective anonymization add complexity. Misuse by malicious entities, public perception issues affecting adoption, and the granularity of data contribute to the privacy landscape. Long-term risks arise from extended data collection, revealing evolving lifestyle patterns. Balancing the benefits of smart meter data with robust security measures, transparent communication, and adherence to privacy regulations is crucial for responsible deployment in smart grid technologies.

In response to these challenges, there is a compelling need to devise innovative strategies for data generation within LV networks, leveraging advanced computational learning techniques. In some instances, the data void is so pronounced that it approaches non-existence.

This research addresses the challenge of smart meters historical data reconstruction using various machine learning techniques, commencing with a limited set of initial data. The study also explores the implications of integrating distributed generation and high-power loads into the LV network, analyzing their effects on predictive models. The significance of this work lies in its potential to enhance the precision and reliability of managing LV networks, facilitating more effective decision-making for network operators.

1.2 Objectives

In the realm of modern energy infrastructure, smart meters in LV networks offer real-time data insights. However, challenges arise in reconstructing historical datasets, especially in data-scarce scenarios. This study aims to pioneer innovative strategies, overcome these hurdles, and enhance the precision and adaptability of historical data reconstruction from smart meter networks.

The primary objectives of this study revolve around the development and evaluation of strategies for reconstructing a synthetic historical dataset derived from the measurement infrastructure within LV networks, specifically smart meters. The overarching goals can be delineated as follows:

Baseline Reconstruction without Additional Information

In the initial phase, the study seeks to establish a baseline for reconstruction. This involves reconstructing a substantial portion of historical data without relying on additional information. Specifically, a predefined set of reference historical data from consumers in a LV network serves as the basis for this reconstruction, providing a foundation for further evaluations. For the cases of new clients where there is any historical data we also intend to validate whether a consumer categorization framework, based on factors such as the type of consumption and production resources, as well as social data like the number of people in the household, allows for the derivation of a synthetic history.

Incorporating Equipment Changes for Improved Reconstruction

Building upon the baseline reconstruction, the study aims to assess the impact of incorporating additional information regarding network equipment changes. For instance, knowledge of a user acquiring an EV or installing PV systems can significantly enhance the accuracy of historical data reconstruction.

Evaluation and Comparison of Synthetic Data Quality

This involves assessing the performance of the reconstruction models by using error metrics to compare the generated data with real data, ensuring accuracy and reliability. Additionally, it includes analyzing the statistical properties (such as mean, variance, skewness, and kurtosis) and physical properties (such as electrical grid sensitivity factors) to ensure that the synthetic data closely matches the characteristics and responses of the original dataset.

Overall, the study focuses on overcoming data scarcity, accurately reconstructing historical data, leveraging additional information for better accuracy, and thoroughly evaluating the quality of the generated data to enhance the completeness and reliability of the reconstructed dataset.

1.3 Document Structure

This document is organized into five main chapters, each serving a distinct purpose and comprehensively exploring various aspects of data-driven decision-making and energy management.

Chapter 1 sets the stage for the study by highlighting the critical need to reconstruct historical data from smart meters. It outlines the primary objectives of this work. The chapter also provides an overview of the document's organization.

Chapter 2 explores the current landscape of smart grids and smart meters, emphasizing their role in modern energy management. It delves into various methodologies for generating synthetic data. The chapter serves as a foundation for understanding the innovative strategies proposed for historical data

Chapter 3 details the research methods used. It describes the generation of synthetic data and the implementation of models for this purpose.

Chapter 4 presents the findings from various reconstruction scenarios. It discusses error management and model comparison. The chapter also examines the impact of incorporating equipment-related changes and compares the statistical and physical properties of the reconstructed data.

Chapter 5 summarizes the study's key findings, highlighting the effectiveness of the proposed machine-learning techniques in reconstructing smart meter data. It also suggests areas for future research and potential improvements in the methodologies discussed.

By following this structured approach, the document provides a clear and comprehensive examination of smart meter data reconstruction, leveraging machine learning techniques to enhance the management of LV networks.

Chapter 2

Background and State of the Art

This chapter is dedicated to providing a contextual framework that is indispensable for thoroughly comprehending the upcoming research. It explores the evolving landscape of data-driven decision-making in energy management. The combination of machine learning methods and advanced data sources is a key and rapidly developing area of study.

2.1 Smart Grids

Smart Grids represent intelligent infrastructures designed for the automatic monitoring of power grids. These grids facilitate bidirectional power and data flows, comprehensive grid description, and the proactive detection, reaction, and prevention of issues within the power grid through the utilization of smart metering techniques. Leveraging digital communication and control systems, smart grids optimize the operation of electrical infrastructure, accommodating the integration of renewable energy sources, energy storage, and electric vehicles. Moreover, these grids possess the capability to automatically adjust in response to changes in demand and power supply.

In contrast, conventional grids primarily respond to consumer demand hierarchically, often lacking significant storage capacity, a notable distinction from smart grids that necessitate energy storage technology for enhanced control of power grid requirements.

The fundamental data source for smart grids, particularly concerning consumption, is derived from smart meters. The use of smart meters, which commenced approximately 50 years ago, marked the inception of smart grid technologies, including automatic metering.

However, the establishment of a Smart Grid entails the integration of advanced technologies to monitor and optimize the electrical infrastructure. Critical components encompass not only the deployment of smart meters for real-time communication but also the incorporation of automation for efficient monitoring and fault detection, as well as the integration of renewable energy sources. Energy storage techniques, demand response mechanisms, and advanced communication networks constitute integral aspects of these grids.

The widespread integration of smart meters culminates in the formation of a Smart Metering Infrastructure, yielding numerous benefits for the Distribution System Operator (DSO) and

consumers alike. For the DSO, real-time data availability facilitates grid monitoring and the prevention of issues such as voltage violations. On the consumer side, the primary advantage lies in the ability to leverage real-time consumption data to optimize energy usage and consequently reduce energy costs. This is particularly evident for consumers with microgeneration systems like PV solar panels, as they can strategically time their electricity usage to align with cost-saving opportunities.

Even more of the Smart Grids advantages are Enhanced electricity transmission efficiency; Expedited electricity restoration post disruptions; Lowered operational and managerial expenses for utilities; Decreased peak demand with bills reflecting actual consumption, devoid of estimations; Augmented assimilation of extensive renewable energy systems; Enhanced grid security [1].

2.1.1 Smart Meters

The metering process started with the rotating disc meters with mechanical displays, the data in these meters was not given automatically, so someone needed to go check the real consumes to the local. Nowadays the metering system is covered mostly with smart meters that report automatically the data.

Smart meters are advanced devices used for measuring and monitoring energy consumption in homes and businesses. These devices are located in every building or service that needs tracking electrical energy consumption. The main advantage of smart meters over conventional meters is their ability to communicate two-way between the consumer and the remote data processing systems.

The implementation of Advanced Metering Systems (AMI) has proven to be highly significant. This integration provides both consumers and system operators with real-time information about electricity consumption. The advantages of AMI integration extend to benefits for both consumers and DSOs, leading to outcomes such as reduced electricity bills for clients and enhanced resources for DSOs. This, in turn, aids in addressing issues like network congestion and voltage violations [1].

2.1.1.1 Advanced Metering Systems

There are numerous benefits and advantages associated with the implementation of the AMI. These include the reduction of costs for meter readings (enabling access to meters that are otherwise challenging to attend to due to their location or security reasons), support for real-time pricing, enhanced fraud detection, and a decrease in the read-to-bill time, to mention a few [2]. As an integral part of the broader smart grid, the AMI can be categorized into three distinct areas: Home Area Network, Wide Area Network, and Utility Companies and Organisations, each with its specific roles and functions, as illustrated in Figure 2.1.

While diverse in their utilities, all these components contribute to common benefits and advantages, including reduced costs, support for real-time pricing, increased fault detection, and reduced

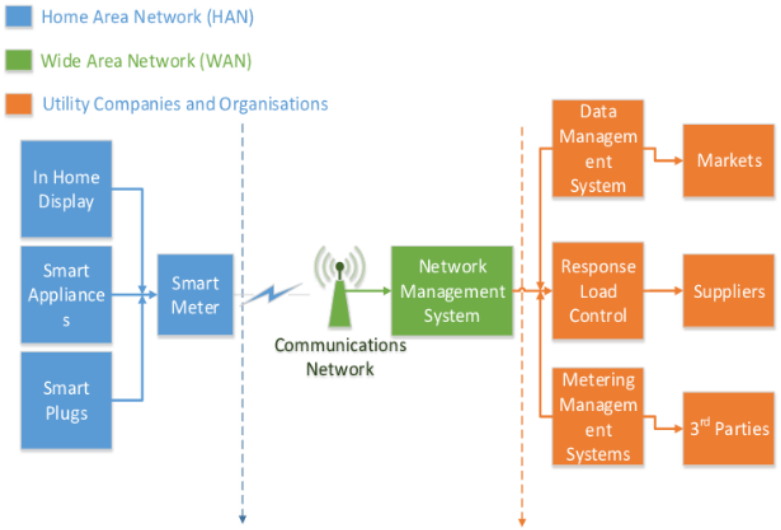


Figure 2.1: Advanced Metering Infrastructure [2]

read-to-bill time. With the addition of more smart meters, the potential for this technology extends beyond aiding consumers in managing their energy to serving as a versatile resource [2].

2.1.2 Consumer's Smart Meters use-cases

Providing User Feedback

A key benefit lies in empowering consumers to align their habits with eco-friendly practices, contributing to broader energy conservation. With the rise of distributed Renewable Energy Sources (RES) like rooftop Photovoltaic (PV) systems, guiding users to synchronize their consumption with local generation becomes pivotal. This shift in approach has reignited interest in energy feedback, especially for prosumers equipped with local energy production setups. The goal is to optimize their grid interaction, not just individually but also within broader contexts such as smart microgrids with Peer-to-Peer (P2P) energy markets. Understanding potential savings and the consequences of decisions becomes crucial for prosumers in such scenarios [3].

An analysis of over 100 studies in a meta-review, as detailed in [4], zeroed in on providing feedback on electricity consumption. The collective findings suggest that feedback has the potential to significantly trim a household's energy use, with reported reductions falling between 5% to 10%. This positive impact is notably enhanced when systems can deliver detailed consumption insights for individual appliances. Another study, [5], evaluates and documents 64 studies, similarly concluding that feedback can realistically contribute to reducing a household's final energy consumption in the range of 5% to 10%.

Recognizing Patterns and Anomalies

Smart meter data also enables the identification of abnormal consumption patterns through signal processing and machine learning techniques. Detecting anomalies in consumption provides an

additional layer of consumer-centric application, although it comes with its set of challenges.

Traditional anomaly detection methods encompass a variety of approaches, including statistical, clustering-based, prediction-based, nearest neighbor, dimensionality reduction, and other complementary models. These approaches often exhibit strong performance and effectiveness within their specific applications. However, despite the effectiveness of traditional anomaly detection methods in certain applications, the need for domain expert intervention and the constraints associated with these methods necessitate further exploration of advanced techniques and frameworks [6]. In the realm of anomaly detection using unsupervised learning algorithms, a substantial body of literature explores the application of deep learning techniques such as predictive models, Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Variational Recurrent Neural Networks (VRNN). For modeling sequential data, especially time series, Recurrent Neural Networks (RNNs) stand out due to their ability to capture long-term temporal dependencies. A study by [7] combined Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) for training a predictive model tailored to energy anomaly detection.

Enabling Demand-Side Flexibility

Demand-side flexibility (DSF) plays a pivotal role in the smart grid, referring to the adjustable share of electricity demand within a specific timeframe. Originally associated with industrial consumers, DSF now extends to domestic and commercial consumers, leveraging controllable appliances and Electric Vehicles (EVs). Aggregating flexibility from various locations, including controllable consumers and RES, contributes significant DSF to the grid. This aggregated flexibility not only benefits a wider consumer base but also offers additional revenue streams, enhances grid stability, reduces environmental impact, and lowers overall generation costs.

In [8], a data-driven scoring approach was used to characterize and rank the users based on the patterns of their interaction with different loads. Several loads were taken into consideration, including EV, dryer, washing machine, dishwasher, as well as AC. The findings indicate that households exhibit significant variations in their potential for demand reduction, and the proposed scoring approach was able to capture these variations. Furthermore, the scoring approach has proven effective in engaging users for targeted and effective demand reduction without a rebound effect. EVs and AC were shown to offer a higher flexibility level than common household appliances.

Load Profiling

While standard load profiles are suitable for many connected customers, they may not capture the unique consumption characteristics of individual consumers. Smart meters address this limitation by providing detailed load profiles, empowering users to understand their energy usage, and making informed comparisons with other households. This increased awareness fosters energy-efficient and responsible behaviors and supports consumer-centric services.

The article in [9] introduces an efficient Non-Intrusive Load Monitoring (NILM)-based Energy Management System (EMS) for residential users on a microgrid. Utilizing Deep Neural Network (DNN) techniques, it accurately extracts appliance-level information from smart meter data, eliminating the need for high-cost sensors. The proposed model analyzes end-users energy consumption behaviors, integrating data into microgrid operations for an optimal and user-centered EMS. Simulation results demonstrate a significant improvement, reducing operational costs and enhancing customer satisfaction by 45% to 65% compared to traditional EMS approaches. The approach offers a cost-effective solution, showcasing its potential for widespread application.

Forecasting Power Demand and Generation

Moreover, smart meters facilitate load forecasting at the individual building level, aiding end customers in managing their energy consumption and correlating current behaviors with future energy costs. Despite the complexity of forecasting at the household level, smart meter data opens numerous opportunities for effective energy management

In reference [10], four machine learning methods—Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Classification and Regression Tree (CART), and LSTM —were employed to forecast both daily consumption peaks and hourly energy consumption in residential buildings based on historical data. It shows that MLPs and LSTM-based approaches demonstrate substantial improvements in short-term (24-hour) demand forecasting, excelling at capturing intricate nonlinear relationships in the data.

Prosumers use forecasting to optimize and manage the resources that generate their energy, like PV or small wind generators. Also, consumption forecasting techniques can help consumers anticipate their future energy needs, so they can plan their local generation and optimize the operation of their Energy Enhancement System (EES) with this information. They can also support the operation of the electricity grid by taking control actions to balance the electricity supply and demand while maximizing self-consumption and profiting from energy arbitrage [3].

2.2 Smart Meters Synthetic Data Generation

The absence of detailed distribution system data poses a significant challenge in advancing data science and machine learning solutions for smart grid applications like load forecasting, dynamic demand response, and behind-the-meter disaggregation. While some efforts have been made to release public datasets, they are constrained by several factors.

Firstly, unlike the transmission level grid operations data regularly published by Independent System Operators (ISOs) and Regional Transmission Operators (RTOs), there is no comparable framework for distribution systems. Consequently, the available distribution system datasets are primarily compiled by individual researchers, leading to limitations in their scope.

Secondly, the scale of available data is insufficient for training sophisticated machine learning models that require large amounts of information. The restricted scope and size of existing public

datasets hinder the application of advanced models that could potentially yield highly accurate results.

Additionally, the privacy of distribution system data obtained from AMI meters is a significant concern. This data often contains Personally Identifiable Information (PII), and complying with regulations requires the implementation of sophisticated algorithms for data anonymization. This privacy concern further restricts the widespread availability of such datasets for research purposes.

Generating synthetic data has become an important approach to overcoming privacy challenges in various contexts, including data from smart meters and other sensitive sources. By using synthetic data, it's possible to create fictional, but reliable, datasets that preserve statistical characteristics and patterns without exposing real user information. In [11], the authors provide a comprehensive review of prevailing data-sharing policies within the electric industry, shedding light on significant limitations while underscoring the potential benefits of increased collaboration. Similarly, the authors of [12] advocate for the open and accessible nature of power system data, emphasizing its pivotal role in facilitating high-quality research and effective policy solutions for the challenges posed by a rapidly evolving grid. Notably, they highlight that the energy sector lags behind other industries concerning data openness. Another crucial consideration is that open access to both data and models helps alleviate redundant efforts and allows for more effective building upon prior research, as emphasized in [12][13][14].

While synthetic grid models are becoming increasingly prevalent as the foundation for numerous research projects,[15] [16] [17], a notable drawback hinders their utility—many of these models typically offer only fixed operating points. Most power system studies and machine learning-based applications heavily depend on large datasets of measurements to comprehensively capture the intricate behaviors of electrical systems. Therefore, the critical capability to map time-series data to a specific grid model is indispensable for enabling the development of novel and enhanced applications. This process ensures the alignment of time-series information with the intricacies of grid models, facilitating the advancement of more accurate and versatile applications in the domain [14].

2.2.1 Stochastic Approaches

In the literature, Markov chains are commonly employed for generating synthetic series (SS) in the analysis of power systems. These models effectively capture time series with statistical properties deviating from the normal distribution [18], [19]. However, concerns have been raised regarding their ability to represent the autocorrelation function (ACF) in series exhibiting seasonal variations. Various modifications have been proposed to address this limitation to enhance ACF representations. Given the significance of ACF and seasonal trends in determining renewable energy supply, the use of Markov chains is typically confined to short-term analyses, where instantaneous power takes precedence over the total energy supplied [18] [19] [20].

In contrast, Regional Transmission Operator models are recognized for generating long-term SS analyses, prioritizing energy considerations over short-term power production by RES [21]. ARIMA models offer improved ACF representations compared to Markov chains [22]. However,

the preprocessing of raw data is necessary to ensure their probability density function adheres to normal distribution [23]. Commonly, numerical derivatives and integral transformations are employed for this purpose, with SS obtained by applying the inverse of the preprocessing process to the model output, such as numerical integration or inverse integral transformation. [20]

While Markov chains and ARIMA models focus on temporal relationships in time series, vector autoregressive (VAR) models are employed for Synthetic Series generation in power systems, emphasizing spatial dependencies among RES time series, especially in expansion planning analyses [20][22][24]. Similar to ARIMA models, VAR models necessitate raw data preprocessing for normal probability density functions. Therefore, integral transformations are often applied for this purpose [24]. VAR models effectively represent statistical features, including ACF, seasonal behavior, and spatial dependencies, making them suitable for generating SS in both long- and short-term power system analyses. However, the raw data must be synchronized to effectively apply the VAR model [20][24].

In [25] a "bottom-up" model is introduced to analyze the load profile of a residential area. The approach involves a two-level synthesis process: first, aggregating individual appliance contributions to the household demand profile, and second, combining various household profiles to determine the overall demand shape of the entire area. To replicate psychological factors influencing residential demand, the model incorporates "behavioral" and "engineering" probability functions. A Monte Carlo extraction process is applied to account for the inherent random variability in demand. Designed with flexibility in mind, the model allows for adjustments to behavioral and engineering functions. This adaptability accommodates changes resulting from technological innovations and evolving socioeconomic factors.

2.2.2 Machine learning Approaches

Machine Learning (ML) emerges as a subset of Artificial Intelligence (AI), encompassing diverse algorithms and mathematical methodologies for problem-solving. In contrast to traditional programming reliant on 'if-then' logic and manual decision schemes, ML algorithms possess the ability to learn through trial and error, autonomously optimizing desired outputs. The essence of ML lies in interpreting existing data, selecting relevant features, and subsequently making predictions for specific tasks. The training phase involves creating the task model, while the testing phase assesses the model's ability to generalize and perform according to its designated purpose. Testing gauges how 'smart' the model has become through its exposure to training. It is crucial to address potential overfitting issues, where the model perfectly fits the training set but struggles to generalize to new datasets. The training process may take the form of supervised learning when the data provides descriptions of situations and outlines the expected behavior for each circumstance. Alternatively, unsupervised learning occurs when the network is exposed solely to unlabeled data without information about anticipated behavior or external influence on the matter [26][27].

In [28] the authors developed a scalable smart meter data generator utilizing the Spark framework. They employed a supervised machine learning approach to create models for generating synthetic data. Additionally, the paper introduced an optimization method that preserves user

groups/cluster information by utilizing clustered seeds. The effectiveness and scalability of the data generator were comprehensively evaluated, and the results demonstrated that the generator can produce scalable smart meter data that closely simulates real-world scenarios. However, their method necessitates the manual creation of features like fluctuations flattening and time series deseasonalizing. Achieving precise modeling of the fundamental causes is a challenging undertaking, as it involves making several assumptions, such as the presence of a Markovian property, which may not always hold. These assumptions can impact the reliability of the synthetically generated data.

In [29] a simulation platform is built on SystemC, a concurrent system-modeling language, to predict household energy consumption with various appliances. The focus is on modeling an advanced metering system using communication-enabled sockets. The simulator's features include accurately predicting power dissipation profiles for tested appliances and offering insights into optimizing energy usage by adjusting the schedule of appliance usage, particularly targeting peak power consumption reduction. The research positions the simulation platform as a valuable tool for designing smart metering systems, allowing realistic testing, verification, and optimization. It suggests the platform's broader applicability, not only for system designers but also for individuals interested in reducing home energy consumption.

In [30] the researchers have developed a GUI-based Synthetic Data Generator using the FBProphet library, focusing on predicting smart meter data—a crucial component of modernized electrical grids. The software adeptly manages substantial data volumes, efficiently producing CSV output files. It automatically eliminates seasonality and trends from the time series source dataset, streamlining algorithm development and minimizing coding efforts. The FBProphet library proves instrumental in addressing challenges associated with time series data, such as handling seasonality and trends. This not only reduces computational complexity and time but also enables scientists, researchers, and developers to test smart meter algorithms securely using synthetic data sets, preserving privacy. The software exhibits potential for expansion by incorporating additional control parameters like seasonal and cluster selection. This prospect for further development holds promise for enhancing the software's performance and versatility. This work holds significance as it offers a practical solution for algorithm testing in data analytics, especially within the realm of smart grids. The utilization of synthetic data sets facilitates efficient and secure testing without the necessity for real-time data collection.

In [31], the authors present an innovative data-driven approach for generating synthetic smart grid data. In contrast to traditional methods involving handcrafted features and assumptions about underlying causes, their approach directly learns the probability distribution of real-time series data. This is achieved through the use of a deep Generative Adversarial Network (GAN) model designed to produce synthetic data that closely mirrors the patterns and characteristics observed in actual smart grid time-series data. The authors applied a probabilistic model and a conditional GAN to generate synthetic smart grid time series datasets. They highlight the uniqueness of their approach, leveraging deep GANs for smart grid applications, and demonstrate the effectiveness of their generated synthetic datasets through statistical tests and classic machine learning tasks.

In [32] a novel time variant Generative Adversarial Network (GAN) solution was developed using the RMSProp optimizer within an existing framework. Trained on energy consumption data from the AEMO website, the model successfully generated synthetic energy consumption data with characteristics closely resembling real-world patterns. The study employed a comprehensive quantitative evaluation, incorporating discriminative score, predictive score, accuracy, precision, and other metrics. Visualization through Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) graphs provided compelling evidence of similar trends between synthetic and real data, affirming the efficacy of the proposed model. The research establishes the Time Variant GAN as a significant advancement, outperforming current frameworks and contributing substantially to the field of energy research. These findings make a substantial contribution to state-of-the-art machine learning and energy analytics. By offering a fresh perspective on synthesizing energy consumption data, the study demonstrates the potential of Time-Variant GANs to accurately replicate dynamic temporal behaviors in diverse datasets. This research not only enhances our understanding of energy consumption patterns but also underscores the broader applicability of advanced machine learning techniques in addressing complex challenges in the field.

In [14] a sophisticated generative framework aimed at overcoming the limitations of fixed operating points in synthetic grid models. Leveraging a vast dataset of phasor measurements exceeding 70 terabytes, the framework employs a combination of GANs, conditional GANs (cGANs), and PCA to capture the dynamic behavior of system loads across different temporal scales. The primary focus on GANs, known for their efficacy in handling non-linear characteristics, allows the framework to generate diverse bus-level load profiles at varying sampling resolutions, ranging from 30Hz to one sample per year and for different time lengths, spanning seconds to a year. The generative scheme encompasses a meticulous four-step process, including data downsampling, independent model training, and subsequent aggregation, offering a flexible and versatile tool for generating synthetic time-series data tailored to specific applications in power system studies and machine learning-based analyses. This research addresses the challenges of fixed operating points in synthetic grid models and contributes to the development of an open-source generative tool. The tool, along with the trained generative models, enables end-users to efficiently generate a wide array of synthetic load data based on selectable parameters. The generative framework proves its adaptability by capturing load behaviours at different timescales, leveraging multiple models to produce realistic time-series data with varying resolutions and time lengths. The potential applications of this generative scheme extend to modelling load scenarios in rapidly evolving electrical grids, accommodating the penetration of electric vehicles and distributed energy resources, and offering a valuable resource for researchers and practitioners in the field.

The artificially generated time series meter data can find applications across utility energy efficiency programs and various building life cycle stages, contributing to energy efficiency enhancement. One significant application involves testing and validating algorithms in the energy domain. Within this context, algorithms are employed to forecast energy savings, pinpoint potential areas for energy retrofit, and assess opportunities for operational improvements in buildings

[33].

Synthetic smart meter data, generated through GANs, serves multiple purposes in preserving privacy and advancing data-driven applications. It offers a solution when real data is limited, enabling the training of machine learning models while adhering to privacy constraints [34].

Beyond training, synthetic data proves valuable in testing and validating data analytics tools and algorithms. Simulating diverse scenarios becomes possible, allowing researchers to explore system responses without compromising individual privacy. This synthetic approach facilitates robustness testing and prepares systems for unforeseen events [34].

Moreover, synthetic data promotes privacy-preserving data sharing for research purposes. By openly distributing synthetic datasets, collaboration and advancements in the field can occur without violating privacy regulations. An additional benefit lies in enhancing anomaly detection algorithms, where synthetic samples contribute to improved model performance [34].

In the context of power grids, synthetic data plays a role in optimizing operations. It aids in developing algorithms for efficient grid management, anomaly detection, and preventive measures. The use of synthetic data faithfully replicates the statistical properties, variations, and outliers present in the original data, making it a reliable substitute for planning system operations without compromising privacy [34].

2.3 Theoretical Background

2.3.1 Artificial Neural Networks

A neural network is a massively parallel process for performing numerical computations. The fundamental element of a neural network is the "neuron," which is a basic processing unit. The concept of the "neuron" as a computational unit is inspired by the biology of neurons in the nervous system, particularly the human nervous system.

In artificial neural networks, "neurons" also have inputs and an output, with non-linear internal processing that transforms the inputs into the output. These neurons are typically organized into layers, with unidirectional connections progressing from the input to the output of the network. The connections between neurons are associated with weights, such that a signal emitted by one neuron is multiplied by the connection weight before entering the next neuron.

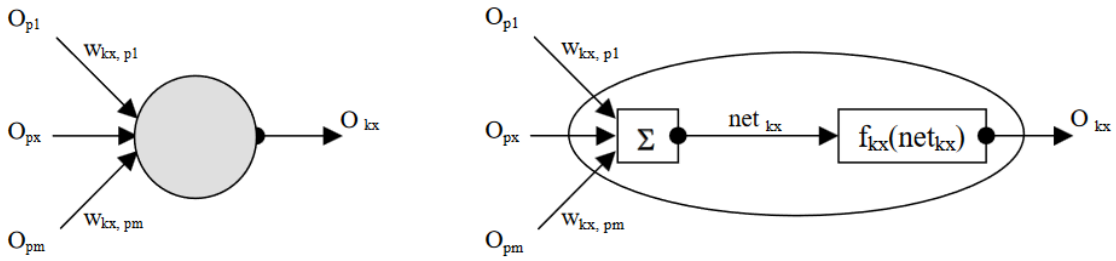


Figure 2.2: Operations executed in each neuron of the Neural Network [26]

A neuron indexed x in layer k receives signals O_{pz} from the neurons in the previous layer p and performs the following transformations (see Figure 2.2) before producing an output signal O_{kx}, which will then activate neurons in the next layer:

1. **Weighted Sum Calculation**: Each input signal O_{pz} is multiplied by its corresponding weight w_{pz}. The neuron sums these weighted inputs along with a bias term b_x:

$$\text{net}_x = \sum_p w_{px} \cdot O_{pz} + b_x \tag{2.1}$$

2. **Activation Function**: The neuron applies a non-linear activation function f to the weighted sum to produce the output signal:

$$O_{kx} = f(\text{net}_x) \tag{2.2}$$

The output O_{kx} is then passed to neurons in the subsequent layer, continuing the process of transformation and propagation through the network.

The function f is the activation function of the neural network. There are numerous types of activation functions, each playing a crucial role in determining the outputs of neurons. The choice of activation function is significant because it greatly affects how the inputs are processed. Specifically, using a nonlinear activation function allows the neural network to model complex, nonlinear systems. Conversely, if a linear activation function is used, the neural network will be limited to modelling only linear systems.

The most commonly adopted activation functions are the linear or ramp function, the step function, and the sigmoid function and its variants. Among these, the ReLU (Rectified Linear Unit) has become extremely popular due to its simplicity and computational efficiency. The ReLU is defined as in 2.3 which means it activates only neurons with positive inputs, helping to avoid the vanishing gradient problem common in activation functions like the sigmoid.

$$f_(x) = \max(0, x) \tag{2.3}$$

However, ReLU has its limitations, such as the "dying neurons" problem, where many neurons can become inactive and stop learning if the inputs are negative. To mitigate this, variants of ReLU were developed, such as Leaky ReLU, which allows a small gradient for negative inputs, and PReLU (Parametric ReLU), where the slope for negative inputs is a trainable parameter.

Recently, the GELU (Gaussian Error Linear Unit) activation function has gained attention for its performance in deep neural networks, especially in natural language models like BERT. The GELU is defined as:

$$f_(x) = x \cdot \Phi(x) \tag{2.4}$$

where $\Phi(x)$ is the cumulative distribution function of the standard Gaussian:

$$\Phi_(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{2.5}$$

The main advantage of GELU over ReLU is that it combines neuron activation to account for Gaussian uncertainty, allowing smoother learning and better performance in some tasks.

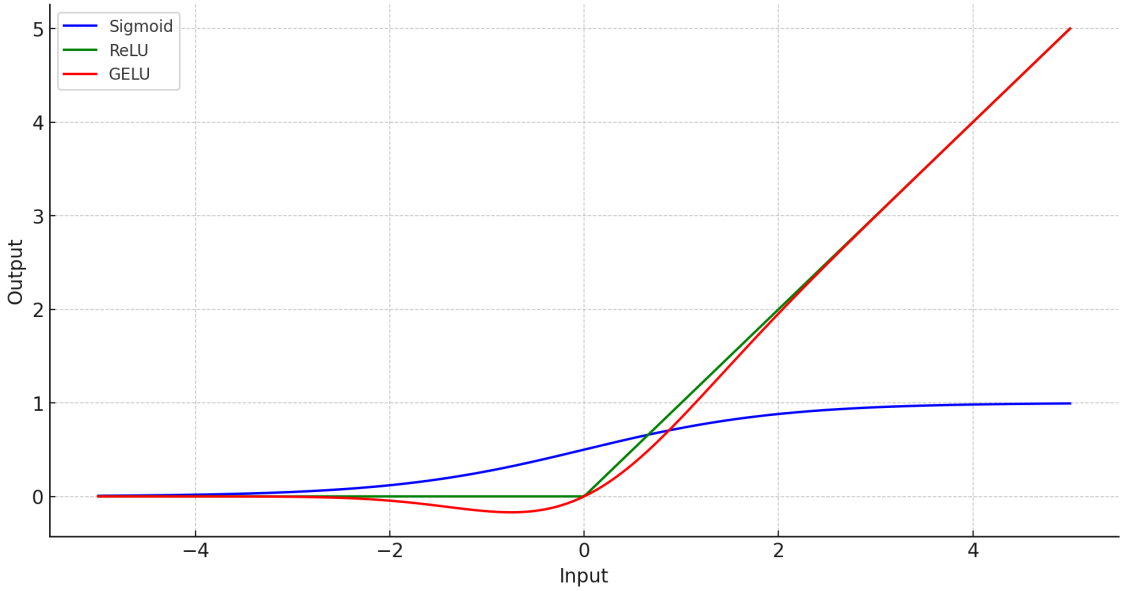


Figure 2.3: Comonly used activation functions

Figure 2.3 ilustrate the sigmoid, ReLU, and GELU activation functions. The blue line represents the sigmoid function, the green line represents the ReLU function, and the red line represents the GELU function.

While Artificial Neural Networks (ANNs) have proven highly effective and sufficient for many real-world tasks, they possess inherent limitations. For instance, conventional ANNs, known as Feedforward Neural Networks (FNNs) (2.4), only facilitate a one-way flow of information. Consequently, they cannot preserve information from prior states, a crucial aspect for addressing numerous problems involving sequential data. To address this limitation and cater to such requirements, RNNs were developed.

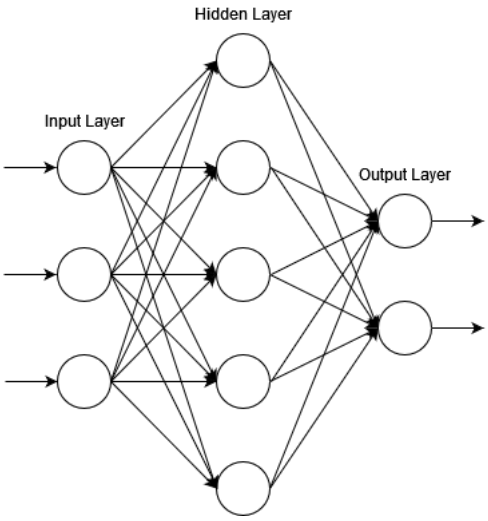


Figure 2.4: Feedforward Neural Network

2.3.2 Recurrent Neural Networks

Recurrent Neural Networks are a type of artificial neural network designed to recognize patterns in sequences of data, such as time series, text, or audio. Unlike feedforward neural networks, which process inputs in a single direction, from the input layer to the output layer, RNNs have connections that form directed cycles, allowing them to exhibit dynamic temporal behavior.

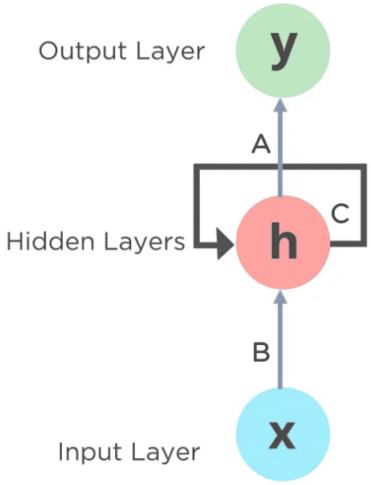


Figure 2.5: Recurrent Neural Network [35]

The key characteristic of RNNs is their ability to maintain a memory of previous inputs through time. This memory makes them particularly effective for tasks where the context of previous inputs is crucial for understanding the current input. For example, in natural language processing tasks like language translation or text generation, RNNs can consider the previous words in a sentence to predict the next word.

However, traditional RNNs suffer from the "vanishing gradient" problem, where gradients diminish exponentially as they are backpropagated through time, leading to difficulties in learning long-range dependencies. To address this issue, various extensions of RNNs have been developed, such as Long Short-Term Memory (LSTM) networks. These architectures incorporate mechanisms to better capture long-term dependencies and alleviate the vanishing gradient problem.

2.3.3 Long Short-Term Memory (LSTM) networks

Long Short-Term Memory (LSTM) is a type of RNN designed to address the vanishing gradient problem that occurs in traditional RNNs. They are specifically designed to prevent the neural network output from either decaying or exploding as it cycles through feedback loops. These feedback loops allow recurrent networks to excel at pattern recognition compared to other neural networks. [36]

LSTM networks present significant advantages over traditional RNNs in two key aspects.

Firstly, LSTMs address the notorious vanishing gradient problem encountered in training traditional RNNs. This issue arises when gradients become exceedingly small during backpropagation, hindering the network's ability to learn from distant past inputs. LSTMs mitigate this problem through their gated architecture, which includes mechanisms to regulate the flow of information and prevent the gradients from vanishing. Consequently, LSTMs exhibit greater robustness in learning long-range dependencies, facilitating more effective training on sequential data. [36]

Secondly, LSTMs excel at capturing long-term dependencies within input sequences. This capability is particularly advantageous in tasks involving complex temporal relationships, such as machine translation, speech recognition, and sentiment analysis. By incorporating memory cells and employing gating mechanisms to control the flow of information, LSTMs can selectively retain relevant information over extended periods. This enables them to capture subtle patterns and relationships across distant time steps, resulting in superior performance on tasks that require modeling intricate temporal dynamics. [36]

LSTMs improve the reliability of long-term memory by incorporating gates. These gates regulate the flow of information by managing self-loops and their weights, enabling the adjustment of the time scale of integration. This means that the LSTM can dynamically adapt to different patterns and sequences of data. You can think of it as a control mechanism that decides which information is important to keep and which can be discarded, allowing the LSTM to remember relevant information over longer periods.

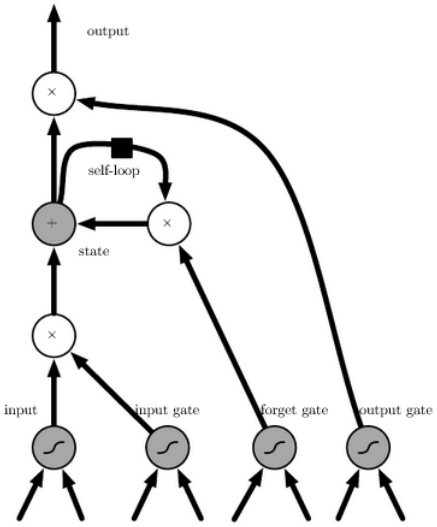


Figure 2.6: Block diagram of the LSTM recurrent network "cell." [37]

The Forget Gate in an LSTM network determines which information to discard. It achieves this by utilizing a sigmoid function to normalize the weights of the self-loops, assigning values between 0 and 1, where 0 signifies deletion and 1 indicates retention of the information as is. This process is mathematically expressed as shown in Equation 2.6, where b represents bias, U denotes input weights, x_t denotes the current input vector, W denotes recurrent weights, and h_t represents the currently hidden layer vector.

$$f_t = \sigma(b_f + U_f x_t + W_f h_{t-1}) \tag{2.6}$$

Similarly, the Input Gate determines the inclusion of new data into the cell. Like the Forget Gate, it employs a sigmoid function:

$$g_t = \sigma(b_g + U_g x_t + W_g h_{t-1}) \tag{2.7}$$

The Output Gate, on the other hand, decides the relevance of information and presents it as an output:

$$q_t = \sigma(b_o + U_o x_t + W_o h_{t-1}) \tag{2.8}$$

Finally, the output of a single cell, h_t, which is the outcome of the computation of the cell state (Equation 2.9) through an activation function (in this case, the hyperbolic tangent) multiplied by the output gate function (Equation 2.8):

$$s_t = f_t s_{t-1} + g_t \sigma(b_s + U_s x_t + W_s h_{t-1}) \tag{2.9}$$

$$h_t = \tanh(s_t) q_t \tag{2.10}$$

Chapter 3

Methodology

This chapter explores the detailed methodology employed in this research to reconstruct absent data for LV network configurations incorporating solar PV and EVs. Our approach aims to address the challenges posed by the scarcity of granular electrical network data and the proprietary nature of critical data, which hinders effective decision-making and network management.

An algorithm was developed to reconstruct the historical data and analyze different scenarios. Various customer consumption profiles were utilized to understand the algorithm's behaviour. The customer voltage profiles were obtained using a radial power flow program developed at INESC TEC, which accurately illustrates the characteristics of LV networks. A one-year database was utilized to analyze error metrics for each scenario and compare the results obtained.

3.1 Data Reconstruction

It employed advanced computational learning techniques, including machine learning algorithms, to reconstruct the historical data. The process includes:

1. **Baseline reconstruction**: Using Linear regression and neural networks to reconstruct baseline data from the initial limited dataset.

2. **Incorporation of equipment-related changes**: Enhancing the accuracy of the reconstructed data by integrating information on changes related to network equipment, such as the addition of photovoltaic systems and electric vehicle chargers.

3. **Statistical Preserved Power Reconstruction (SPPR)**: Trying to preserve the statistical properties of the injected power time series, prioritising it over deterministic error metrics.

4. **Error metrics evaluation**: Evaluating the performance of the data generation process by using error metrics to compare the obtained data against the real data, ensuring that it is accurate and reliable.

5. **Statistical properties comparison**: Comparing the statistical properties of the obtained data with those of the original dataset to ensure that the it retains similar characteristics, such as mean, variance, skewness, and kurtosis.

6. **Physical properties analysis**: Comparing the physical properties of the estimated and real data, particularly focusing on the electrical grid's sensitivity factors. This involves ensuring that the data accurately reflects the response of the electrical grid to various changes and stresses. The sensitivity coefficients of LV networks have been calculated using a programme developed at INESC TEC. [38]

3.2 Implemented algorithms

The baseline reconstruction is divided into two main categories: the reconstruction of consumer voltage profiles and active power consumption profiles. For the base reconstruction of both groups, linear regression models were employed, followed by models utilising Recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM's).

We evaluate the methods under three scenarios to assess the impact of data availability:

- **Single Meter Data:** Reconstruction of consumption data for one smart meter.
- **Single Phase Data:** Reconstruction of consumption data for all the meters in the same phase.
- **Full Grid Data:** Reconstruction of consumption data for all the meters of the grid.

3.2.1 Baseline time series Reconstruction

The first part of this work was dedicated to reconstructing time series for LV consumers, beginning with a small data set and trying to reconstruct the major part of the dataset without any extra consumer data. To achieve that, we use two different approaches:

3.2.1.1 Linear Regression

Linear regression models the relationship between voltage and active power y and n input variables $x_1, x_2, \ldots, x_n$. The model finds the best-fit linear equation:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{3.1}$$

where $a_0, a_1, \ldots, a_n$ are coefficients determined from the data.

3.2.2 Recurrent Neural Network (RNN)

RNNs are suited for sequential data and can handle multiple input variables. They use a structure that allows the network to retain information over time. For n inputs at each time step t, the RNN update can be described as:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t) \tag{3.2}$$

where h_t is the hidden state at time t, h_{t-1} is the hidden state at the previous time step, x_t is the input vector of size n, and σ is an activation function.

Long Short-Term Memory (LSTM) network architectures are a specific kind of RNN, and they incorporate mechanisms to capture long-term dependencies better and alleviate the vanishing gradient problem. They are the ones that will be used in this work.

3.2.3 Incorporation of Equipment-Related Information

The method aims to improve baseline data reconstruction by incorporating additional information about changes in network equipment. For example, knowledge of a user acquiring an EV or installing PVs can significantly enhance the accuracy of historical data reconstruction.

In order to incorporate the consumption data associated with EV chargers observed during the initial two months into the model, we proceeded to replicate this consumption pattern over the subsequent ten months of the dataset. In doing so, we considered the charging hours of each bus and the maximum charging power.

On the other hand, the information about PV generation was obtained through a machine learning model involving neural networks, considering the irradiance of the provided location [39].

3.2.4 Statistical Preserved Power Reconstruction (SPPR)

This approach aims to improve the statistical properties of the injected power time series. The approach considered for this point is recurrent neural networks with some modifications to those used in the basic reconstruction. The main change is the loss function used to train the network. The loss function used in the basic reconstruction was the mean squared error, which only tries to minimize the distance between real and estimated values without guaranteeing that the estimated time series follows the same probability distribution. In this case, a combination of two functions was used: a loss function based on quantile loss in combination with mean squared error.

Quantile Loss is a loss function used in regression models to estimate specific quantiles of the conditional distribution of the data (e.g., the 90th percentile or 50th percentile if we want the median).

$$\text{Quantile Loss}(\hat{y}, y) = \begin{cases} (1-\tau)\cdot(y-\hat{y}) & \text{if } y \geq \hat{y} \\ \tau\cdot(\hat{y}-y) & \text{if } y < \hat{y} \end{cases} \tag{3.3}$$

where:

- $\hat{y}$ is the model's estimation,
- y is the true value,

- τ is the desired quantile (for example, $\tau = 0.9$ for the 90th percentile).

This function penalizes errors based on how they relate to the quantile τ, allowing the model to focus on estimating different parts of the data distribution.

It is feasible to combine different loss functions to capture multiple properties of the data. This is typically achieved using a weighted sum of the individual loss functions. For instance, to balance deterministic properties (captured by Mean Squared Error, MSE) with statistical properties (captured by Quantile Loss), we can define a combined loss function as follows:

$$\text{Loss}_{\text{combined}} = \alpha \cdot \text{MSE} + \beta \cdot \text{Quantile Loss}$$

In this formulation, α and β are coefficients that determine the relative importance of each component of the loss function. The coefficient α adjusts the weight of the MSE, which focuses on the estimation accuracy, while β adjusts the weight of the Quantile Loss, which aims to capture specific statistical properties of the data.

3.3 Features and targets

Considering the concepts discussed in the previous section, we established the input features for the model. Chronological variables, in particular, require special attention. These variables are simply sequences of numbers repeated multiple times throughout the dataset. For instance, the hour of the day ranges from 0 to 23 and repeats across the data. Similarly, the days of the week are represented by numbers 0 through 6 (with 0 representing Sunday and 6 representing Saturday)

However, using these variables in their raw form can lead to significant misinterpretations. For example, when the model processes the hour of the day, it might interpret the values 0 and 23 as being far apart, with a difference of -22 hours, instead of the actual 1-hour difference. This misinterpretation can lead to incorrect conclusions.

To accurately capture the cyclical behaviour of our data, we used Fourier Transformations to represent the week's cyclical behaviour and the hour's cyclical behaviour.

The transformation of these variables into cyclical form ensures that their periodic nature is preserved and correctly interpreted by the model. Specifically, we applied the following Fourier transformations to convert these variables into cyclical variables:

$$x = \sin\left(\frac{f \cdot 2\pi}{\max(f)}\right) \tag{3.4}$$

$$y = \cos\left(\frac{f \cdot 2\pi}{\max(f)}\right) \tag{3.5}$$

where f is the original chronological variable (such as the hour of the day or the hour of the week), and $\max(f)$ is the maximum value of f in its cycle (23 for hours of the day, 7 for days of the week).

In our case, the equations are the following:

$$Hour_s = \sin\left(\frac{Hour \cdot 2\pi}{24}\right) \tag{3.6}$$

$$Hour_c = \cos\left(\frac{Hour \cdot 2\pi}{24}\right) \tag{3.7}$$

$$Week_Day_s = \sin\left(\frac{Weekday \cdot 2\pi}{7}\right) \tag{3.8}$$

$$Week_Day_c = \cos\left(\frac{Weekday \cdot 2\pi}{7}\right) \tag{3.9}$$

By applying these transformations, we can accurately capture and represent the cyclical nature of the chronological variables in our dataset, ensuring that the model can learn and understand the inherent periodic patterns.

In addition to cyclical transformations, models handling time-series data utilize variable lags at intervals like 24, 48, 168, and 336 hours. These lags allow the model to capture short-term temporal dependencies (daily fluctuations) and long-term patterns (weekly or seasonal trends).

The complete set of variables that were used is presented in table 3.1:

Variable	**Description**
$V_i(t-24h)$	Voltage magnitude at bus i at the same hour 1 day before period t
$Vi_(t-48h)$	Voltage magnitude at bus i at the same hour 2 days before period t
$V_i(t-168h)$	Voltage magnitude at bus i at the same hour 7 days before period t (1 week)
$V_i(t-336h)$	Voltage magnitude at bus i at the same hour 14 days before period t (2 weeks)
$V_x_PT(t)$	Voltage magnitude at the PT in phase x at period t
Sin(WD)	first cyclic variable that represents the day of the week
Cos(WD)	first cyclic variable that represents the day of the week
Sin(H)	first cyclic variable that represents the time at each moment
Cos(H)	second cyclic variable that represents the time at each moment
$Pinj_i(t-30min)$	Power at bus i 30 min before period t
$Pinj_i(t-24h)$	Active power at bus i at the same hour 1 day before period t
$Pinj_i(t-48h)$	Active power at bus i at the same hour 2 days before period t
$Pinj_i(t-168h)$	Active power at bus i at the same hour 7 days before period t (1 week)
$Pinj_i(t-336h)$	Active power at bus i at the same hour 14 days before period t (2 weeks)
$Pinj_x_PT(t)$	Active power at the PT in phase x at period t

Table 3.1: Features and their descriptions

3.4 Train, Validation and Test Split

In our study, we utilized different data-splitting strategies for linear regression and neural networks to ensure optimal performance for each model type.

Linear Regression

For the linear regression model, we adopted a split where 20% of the data was used for training, and the remaining 80% was allocated for testing. This approach was chosen to provide the model with a substantial amount of data for evaluating its performance, ensuring that it generalizes well to unseen data. The larger test set helps in assessing the robustness of the linear regression model, which is particularly sensitive to overfitting when the training data is not representative of the overall data distribution.

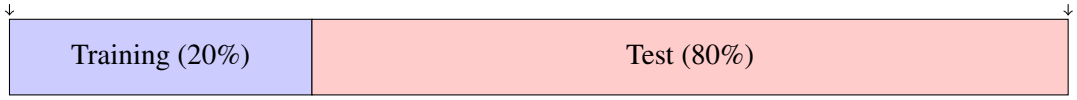


Figure 3.1: Linear Regression Data Split

LSTM

In contrast, we implemented a different strategy for the LSTM models. The data was divided into 10% for training, 10% for validation, and 80% for testing. The rationale behind this split is rooted in LSTM's complexity and data-hungry nature. A smaller training set encourages the model to learn efficiently from a limited dataset, while the validation set is used to tune hyperparameters and prevent overfitting during the training process. The extensive test set, similar to the linear regression approach, ensures a thorough evaluation of the model's generalization capabilities.

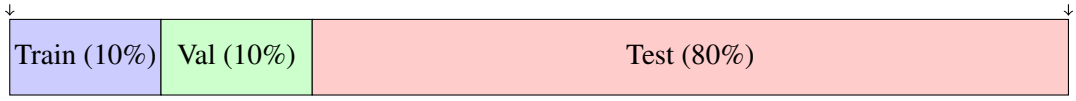


Figure 3.2: Neural Network Data Split

Chapter 4

Results

This chapter presents the outcomes of this study's various methodologies and models. It encompasses the performance metrics, visualizations, and comparative analysis of the data reconstruction techniques applied to the smart meter datasets. The results highlight the efficacy of the baseline reconstructions and the enhancements achieved through incorporating equipment changes. Each subsection will detail the quantitative and qualitative insights derived from the experimental setups, providing a comprehensive understanding of the model's performance and impact on LV network management.

4.1 Data overview

The initial phase of our methodology focuses on data collecting and processing, a critical step to ensure the accuracy and reliability of the generated data. This involves the following steps:

1. **Data Collection**: Gathering available historical data from smart meters and other relevant sources.
2. **Data Transformation**: Obtained voltage historical data using the gathered power injection data and a radial power flow program developed in INESC TEC.

4.1.1 Data Collection

The grid model collected for this project includes the widely used model that accurately represents a Portuguese radial grid. These radial grid models provide a detailed representation of electrical grid topology, serving as a foundational element for developing and testing the proposed tool.

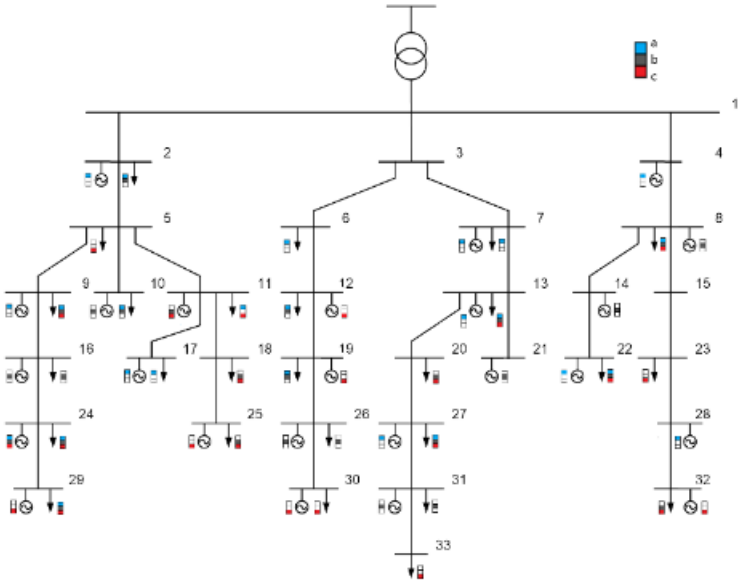


Figure 4.1: Traditional Portuguese Grid [1]

The time series for electricity consumption originated from a smart metering trial overseen by the Commission for Energy Regulation (CER) in Ireland [40]. The PV generation data was gathered in Évora, Portugal, as part of the SuSTAINABLE project [39]. The EV charging data was collected from a study by the Scottish and Southern Energy Power Distribution [41].

4.1.2 Data Transformation

For this work, we use a 1-year consumption dataset. With this dataset, we employ an unbalanced backwards/forward sweep power flow developed at INESC TEC to generate the voltages using the grid model and the consumption dataset. The main challenge here is defining the transformer bus's reference voltage. Four different methods were considered:

- **Method 0** - Fix the bus at 230 V: This is the simplest model, where we fix the LV transformer bus at 230 V and run the power flow.
- **Method 1** (Figure 4.2) - Use the relationship between the full power injected at that bus and the voltage at that bus.
- **Method 2** - Fix the bus at 230 V and apply the transformer admittance to try to represent the losses at the transformer.
- **Method 3** (Figure 4.3) - A mix of Method 1 and Method 2: We use the relationship to obtain the reference voltage and then apply the transformer admittance.

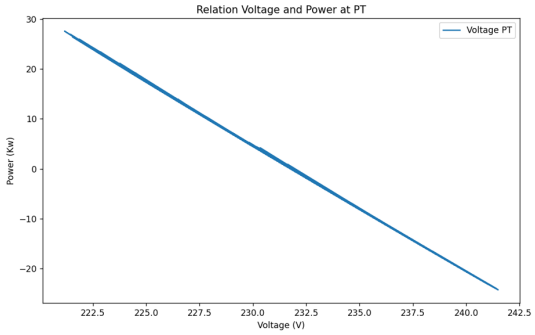


Figure 4.2: Method 1

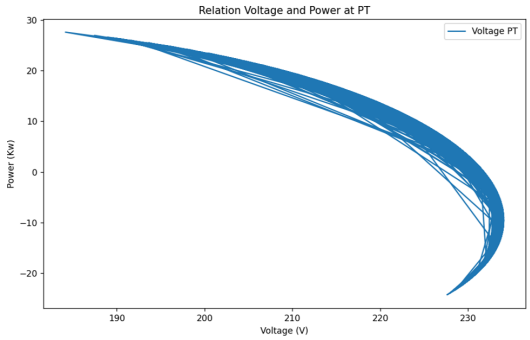


Figure 4.3: Method 3

The original method developed does not fully represent the behaviour of the LV Grid due to insufficient information about the transformer station. To address this issue, an imaginary line with the reactance value of the transformer was added to the grid model to represent the transformer's admittance. The voltage at the "Medium Voltage bus" was then determined using the linear relationship between the power consumption on the grid and the imaginary line representing the transformer connected to the "LV bus." Consequently, we implemented Method 2 and Method 3, which provide a more accurate description of the behaviour of a LV radial network and will be used from now on.

4.2 Evaluation metrics

In evaluating the performance of our models on the test dataset, we considered three primary error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics provide valuable insights into the accuracy and precision of our models' generated values. Specifically, MAE measures the average magnitude of errors between generated and actual values, while RMSE penalizes larger errors more heavily. Additionally, R^2 quantifies the proportion of variance in the target variable explained by the model. Each metric offers a unique perspective on model performance, allowing us to assess the generated value's accuracy and overall model fit.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\text{real}_i - \text{generated}_i| \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{real}_i - \text{generated}_i)^2} \tag{4.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\text{real}_i - \text{generated}_i)^2}{\sum_{i=1}^{n}(\text{real}_i - \bar{\text{real}})^2} \tag{4.3}$$

Where:

- n is the number of observations.

- real_i is the actual value of the target variable for the ith observation.
- generated_i is the generated value of the target variable for the ith observation.
- $\bar{\text{real}}$ is the mean of the actual values of the target variable.

4.3 Probability distribution properties

In the analysis of both real and generated data, we consider various statistical properties to ensure the generated data accurately reflects the characteristics of the real data. These properties include:

Mean

The mean (μ) is a measure of central tendency that represents the average value of the dataset. It is calculated by summing all the values and dividing by the number of values.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.4}$$

where N is the number of data points and x_i is the value of the i-th data point.

Standard Deviation

The standard deviation (σ) measures the dispersion or variability of the data around the mean. It is calculated as the square root of the variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4.5}$$

where μ is the mean of the data.

Kurtosis

Kurtosis (K) describes the "tailedness" of the data distribution. It is calculated as:

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^4 - 3 \tag{4.6}$$

where μ is the mean and σ is the standard deviation of the data. The value 3 is subtracted to make a normal distribution's kurtosis equal zero.

Skewness

Skewness (S) measures the asymmetry of the data distribution. It is calculated as:

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \tag{4.7}$$

where μ is the mean and σ is the standard deviation of the data.

Correlation

The correlation coefficient (ρ_{ij}) between two variables x_i and x_j is calculated as:

$$\rho_{ij} = \frac{\text{cov}(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} \tag{4.8}$$

where $\text{cov}(x_i, x_j)$ is the covariance between x_i and x_j, and σ_{x_i} and σ_{x_j} are the standard deviations of x_i and x_j, respectively.

Probability Density Function (PDF)

The PDF $f(x)$ describes the likelihood of different outcomes in a continuous random variable. For a normal distribution, it is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.9}$$

where μ is the mean and σ is the standard deviation.

Autocorrelation Function (ACF)

The ACF $r(k)$ measures the correlation of a time series with its own past values. It is calculated as:

$$r(k) = \frac{\sum_{t=1}^{N-k}(x_t - \mu)(x_{t+k} - \mu)}{\sum_{t=1}^{N}(x_t - \mu)^2} \tag{4.10}$$

where k is the lag, N is the number of data points, and μ is the mean of the series.

Partial Autocorrelation Function (PACF)

The PACF $\phi(k)$ measures the correlation between a time series and its lagged values, controlling for the values of the time series at all shorter lags. It is calculated by solving the Yule-Walker equations, often using statistical software.

Seasonality, Residuals, and Trend

Decomposing a time series into seasonality, residuals, and trends can be represented as:

$$x_t = T_t + S_t + R_t \tag{4.11}$$

where x_t is the observed value at time t, T_t is the trend component, S_t is the seasonal component, and R_t is the residual (or irregular) component. This decomposition helps in understanding the underlying structure of the time series.

By thoroughly analyzing these statistical properties, we ensure that the generated data maintains the characteristics of the real data, enabling its use in various analytical and modelling applications.

4.4 Voltage Time Series Reconstruction

This section unveils the findings and insights from the voltage time series data reconstruction.

4.4.1 Single-meter Reconstruction

For the reconstruction of single-meter measurements using Linear Regression, focusing on bus 33 of phase A of the grid presented in 4.1, four models were developed using Linear Regression. LR_0 was only used for this approach so the patterns of the voltage-time series data can be understood.

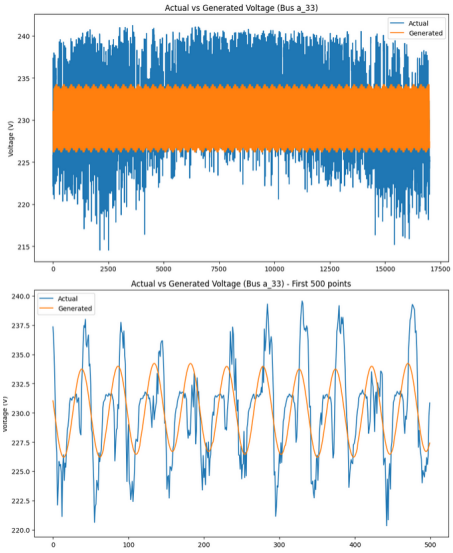
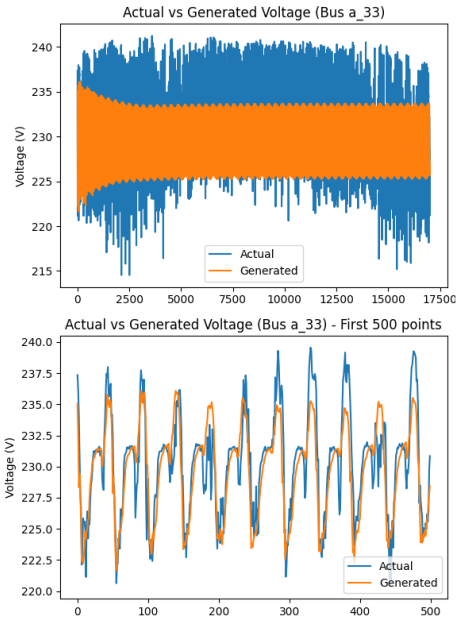
Model	Variables	Graphics	Comments
LR_0	Sin(WD) Cos(WD) Sin(H) Cos(H)		As expected, for this bus, this model can only capture the time series' daily patterns but cannot catch the voltage peaks, this model is only presented here to demonstrate the cyclic behaviour of the voltage time series, and it won't be replicated.
LR_1	$V_{33}(t-24h)$ $V_{33}(t-48h)$ $V_{33}(t-168h)$ $V_{33}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H)		This model, unlike the previous one, captures some peaks at the beginning of the test set due to its reliance on past measurements. However, as the reconstruction progresses, the model's ability to retain information diminishes, leading to a point where the result becomes very similar to that of model 0.

Table 4.1: Comparison of LR Models for Single-meter Voltage Reconstruction (Part 1)

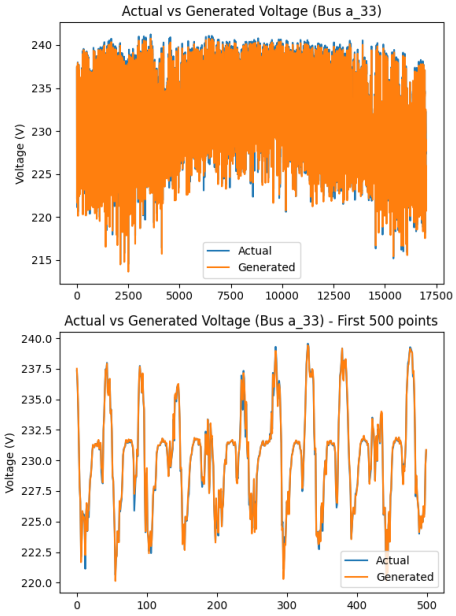
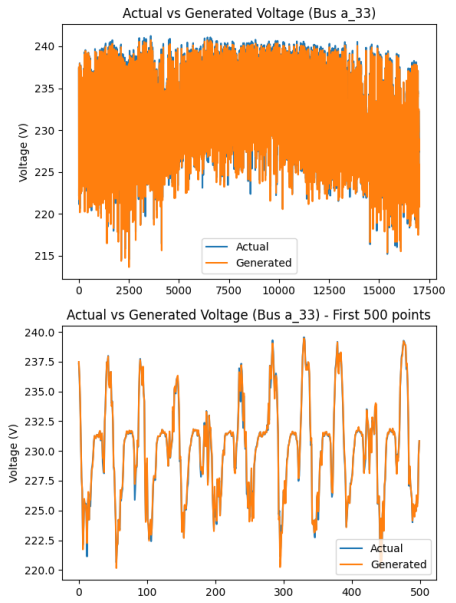
Model	Variables	Graphics	Comments
LR_2	$V_{33}(t-24h)$ $V_{33}(t-48h)$ $V_{33}(t-168h)$ $V_{33}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		Including the Secondary Substation (SS) real-time measurement in the model significantly enhances the results, as expected, as it provides nearly all the information about the voltage profile in the bus, even when it is far from the meter where the voltage is being monitored.
LR_3	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		This model was introduced to understand the impact of using only the SS voltage and cyclic variables and to assess whether a simpler model can yield good results for this bus.

Table 4.2: Comparison of LR Models for Single-meter Voltage Reconstruction (Part 2)

For the reconstruction of single-meter measurements using LSTM, focusing on bus 33 of phase A of the grid presented in 4.1, three models were developed using LSTM's:

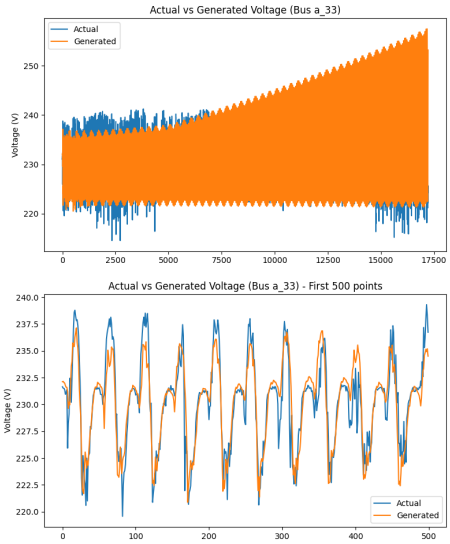
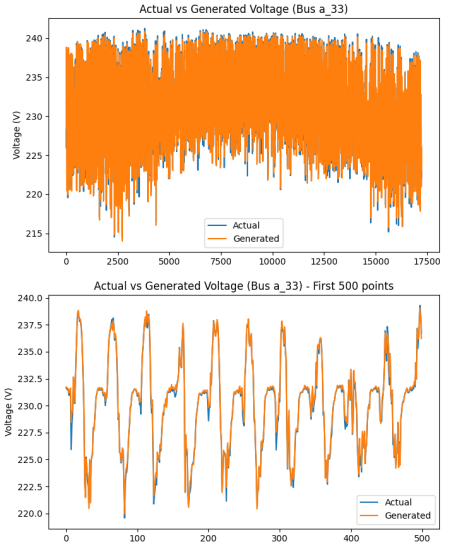
Model	Variables	Graphics	Comments
NN_1	$V_{33}(t-24h)$ $V_{33}(t-48h)$ $V_{33}(t-168h)$ $V_{33}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H)		This model retains more information than his equivalent in linear regression, as was expected; from this, it is possible to conclude that until the 3000 timestamps, nearly two months, it is possible to reconstruct this bus without any real-time information.
NN_2	$V_{33}(t-24h)$ $V_{33}(t-48h)$ $V_{33}(t-168h)$ $V_{33}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		As in linear regression, adding the SS voltage significantly enhances the reconstruction model, as expected, as it closely mirrors this bus behaviour.

Table 4.3: Comparison of LSTM Models for Single-meter Voltage Reconstruction (Part 1)

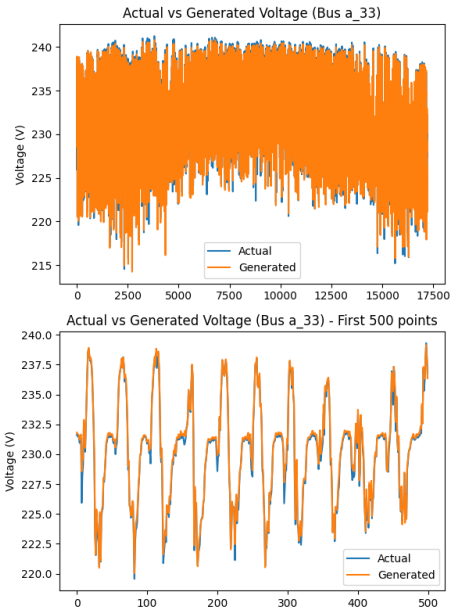
Model	Variables	Graphics	Comments
NN_3	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		As in linear regression, only using the Fourier transformations and the SS real-time measurements get better performance for this bus then the previous approach.

Table 4.4: Comparison of LSTM Models for Single-meter Voltage Reconstruction (Part 2)

To quantitatively assess the most effective model in reconstructing stress history, error metrics were calculated as detailed in Section 4.2. The following results were obtained for LR models and LSTM models.

Model	Elapsed time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score	Max Absolute Error
LR_0	142.812	6.778	5.171	0.373	24.892
LR_1	131.904	7.691	5.553	0.197	28.217
LR_2	142.746	0.295	0.192	0.999	2.714
LR_3	147.104	0.299	0.194	0.999	2.698
NN_1	1305.765	8.485	6.215	-2.957	28.898
NN_2	1385.750	0.749	0.688	0.992	3.453
NN_3	207.733	0.311	0.212	0.994	2.732

Table 4.5: Error metrics for voltage time series reconstruction (single-meter)

LR_0 has relatively high RMSE and MAE values, indicating a significant difference between the generated and actual voltage. Its R-squared score is also low, suggesting a weak correlation between the generated and actual values. This model likely only captures the daily voltage cycle, not the voltage peaks.

LR_1 and NN_1 show some improvement over basic models in terms of RMSE and MAE, suggesting a slightly better ability to capture voltage variations. However, their R-squared scores remain low, indicating limitations in capturing the complex voltage patterns. The keynote for this model, in particular model NN_1, is its capability to keep a lot of information for nearly two months; this could lead to an essential way to reconstruct data without any real-time measurement.

One way to deal with this is to retrain the model every 2/3 months with one month of data; this could be one alternative; the other is to use real-time voltage at the transformer station. It will always depend on the customer's willingness to pay in data terms.

Models incorporating real-time voltage data from the substation (LR_2, NN_2, LR_3, NN_3) exhibit significantly lower RMSE and MAE values than others. This demonstrates a substantial improvement in accurately reconstructing the voltage profile. Their R-squared scores are close to 1, indicating a strong correlation between the generated and actual voltage values.

The analysis of error metrics (RMSE, MAE, R-squared Score) revealed that models incorporating real-time substation voltage data significantly outperform others in reconstructing voltage time series. Linear Regression models (LR_2 and LR_3) that leverage this real-time data achieve both high accuracy (low RMSE and MAE) and strong correlation with actual voltage (high R-squared Score). Additionally, they are likely computationally faster than their LSTM counterparts (NN_2 and NN_3), which achieve similar accuracy. Therefore, prioritising accuracy and efficiency for this single-meter voltage time series reconstruction scenario leads to LR_2 or LR_3 (Linear Regression with real-time substation voltage) as the best model choices.

4.4.2 Single-phase Reconstruction

This subsection will focus on the historical reconstruction with Linear Regression for an entire phase and compare the total phase error for each type of approach. In this case, only variants of the best models from the previous case were considered: LR_2 and LR_3, as these stood out in performance for the reconstruction of just one meter.

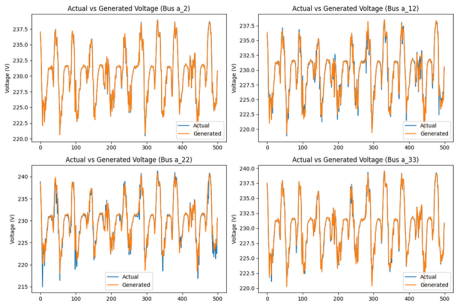
Model	**Variables**	**Graphics**	**Comments**
LR_1	For each bus i in phase a: $Vi(t-24h)$ $Vi(t-48h)$ $Vi(t-168h)$ $Vi(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		This model can explain really well the voltage profiles for all the smart meters of phase a, due to the inclusion of the substation voltage a real-time. This inclusion significantly enriches the model's accuracy and robustness.

Table 4.6: Details of Model 1 for Single-phase Voltage Reconstruction

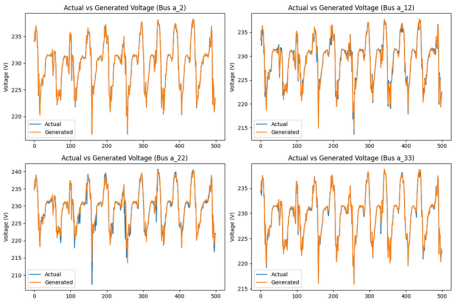
Model	Variables	Graphics	Comments
LR_2	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		Model 2 offers a more straightforward approach compared to Model 1. It utilizes readily available substation voltage data, along with harmonic information (Sin(W)/Sin(H) and Cos(W)/Cos(H), to achieve voltage reconstruction. This reduces the number of historical voltage measurements required, potentially lowering the computational burden while maintaining acceptable accuracy by including key influencing factors.

Table 4.7: Details of Model 2 for Single-phase Voltage Reconstruction

This subsection describes the results of using an LSTM model based on the NN_3 model for single-meter reconstruction. Only this model was used because it achieved the best results for the LR single-meter reconstruction, and we tried to generalize it for the entire phase.

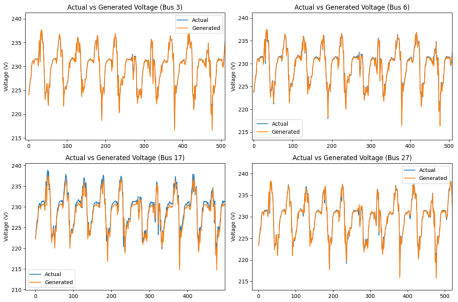
Model	Variables	Graphics	Comments
NN_1	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT}		As observed in its analogous LR model, this LSTM model similarly demonstrates the capacity to elucidate the voltage profiles for all phases of a smart meter, largely attributable to the incorporation of substation voltage in real-time.

Table 4.8: Comparison of LSTM Models for Single-phase Voltage Reconstruction

The provided table (Table 4.9) summarizes the performance of three single-phase voltage reconstruction models: LR_1, LR_2, and NN_1.

Model	Elapsed time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score
LR_1	145.429	0.443	0.237	0.998
LR_2	126.343	0.436	0.235	0.998
NN_1	169.791	0.513	0.324	0.998

Table 4.9: Error Metrics for voltage time series reconstruction (single-phase)

LR_2 is the most favourable model for single-phase voltage reconstruction based on the error metrics. It offers the fastest execution time while maintaining comparable or slightly better accuracy than the other models. However, the performance differences are relatively minor. NN_1 might be a less attractive option due to its higher computational cost.

4.4.3 Full Grid Reconstruction

This subsection will focus on presenting the results related to reconstructing the voltage time series across the entire network. For this purpose, only a linear regression model was used, as it yielded the best results for reconstructing a single phase (model LR_2). The variables used were:

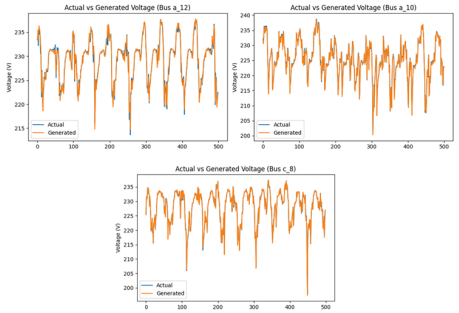
Model	Variables	Graphics	Comments
LR_1	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} V_{b_PT} V_{c_PT}		This figure present the historical reconstruction results across three different phases with Linear Regression reconstruction, allowing us to understand the effect of phase differences on historical reconstruction.

Table 4.10: Comparison of LR Models for Full Grid Voltage Reconstruction

Continuing the previous procedure for the historical voltage reconstruction of a single phase, only one model was used for the voltage reconstruction of the entire network using LSTM, utilizing the following variables:

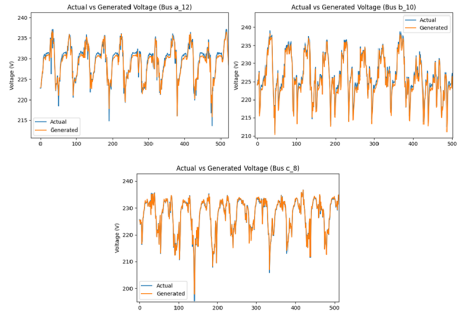
Model	Variables	Graphics	Comments
NN_1	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} V_{b_PT} V_{c_PT}		This figure present the historical reconstruction results across three different phases with LSTM reconstruction, allowing us to understand the effect of phase differences on historical reconstruction.

Table 4.11: Comparison of LSTM Models for Full Grid Voltage Reconstruction

The provided table (Table 4.12) summarizes the performance of two Full Grid voltage reconstruction models: LR_1 and NN_1.

Model	Elapsed time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score
LR_1	176.554	0.653	0.360	0.993
NN_1	196.339	0.850	0.511	0.988

Table 4.12: Error Metrics for voltage time series reconstruction (Full Grid)

Based on the error metrics provided in Table 4.12, the LR model (LR_1) outperforms the LSTM model (NN_1) in terms of all three metrics. LR_1 has a lower RMSE and MAE, indicating that it produces reconstructions that are closer to the original voltage time series on average. Additionally, LR_1 has a higher R-squared score, which implies that it explains a more significant proportion of the voltage-time series variance than NN_1.

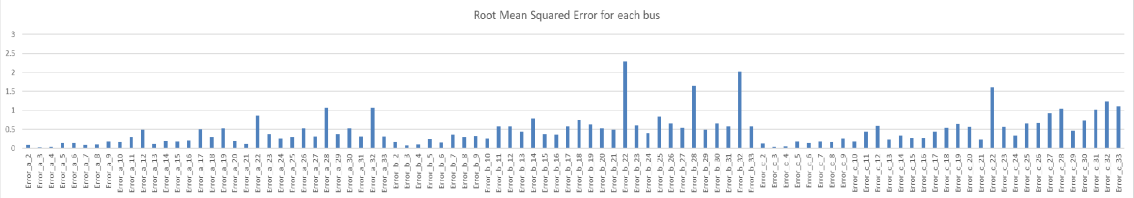


Figure 4.4: Error for each bus using LR forVoltage reconstruction

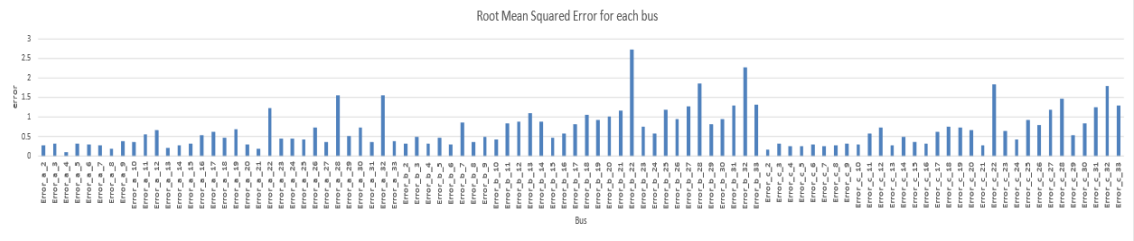


Figure 4.5: Error for each bus using LSTMs for Voltage reconstruction

As shown in Figure 4.4 and 4.5, the error increases as we move further from the SS, which is expected since the influence of the Secondary Substation voltage on the historical reconstruction diminishes. It is also worth noting that the highest error peaks occur at buses 22, 28, and 33. This happens because these nodes are located at the extremities, or close to it in the case of bus 32, of the network, as illustrated in Figure 4.1.

The behaviour of the voltage reconstruction of the a_33 bus over the three approaches was analysed. It was found that the results were almost the same when linear regressions without recursive variables were used. Similarly, when LSTMs were also used without recursive variables, there was only a small variation. This variation was found to be due to the random initialisation of the weights. The behaviour of the aforementioned bus is analysed when recursive variables are employed in the context of linear regressions. It is observed that a slight discrepancy exists between the single-meter reconstruction and the single-phase reconstruction, with the former exhibiting a superior outcome. This is attributed to the fact that in the single-phase reconstruction,

lags from all the meters in that phase are utilised as input, which may introduce some noise when reconstructing the a_33 node.

4.5 Power Injection Time series Reconstruction

This section presents the results of the injected power time series reconstruction for LV customers. Unlike the voltage profiles of an LV consumer, which have clear daily and seasonal patterns, these patterns are not as visible in the context of injected power. This makes the historical reconstruction of a consumer's power profile more challenging.

4.5.1 Single-meter Reconstruction

Three models were implemented using LR to reconstruct bus 29 Power Injection historical.

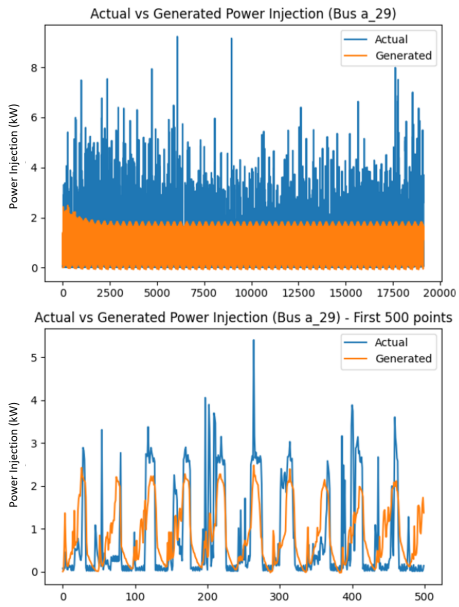
Model	Variables	Graphics	Comments
LR_1	$Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H)		It is evident that the model can accurately explain the initial days of the dataset. However, after a certain point, it completely loses the ability to reconstruct the desired history. This occurs due to the successive propagation of error, resulting in only the cyclic variables having an effect.

Table 4.13: Comparison of LR Models for Single-meter Active Power Reconstruction (Part 1)

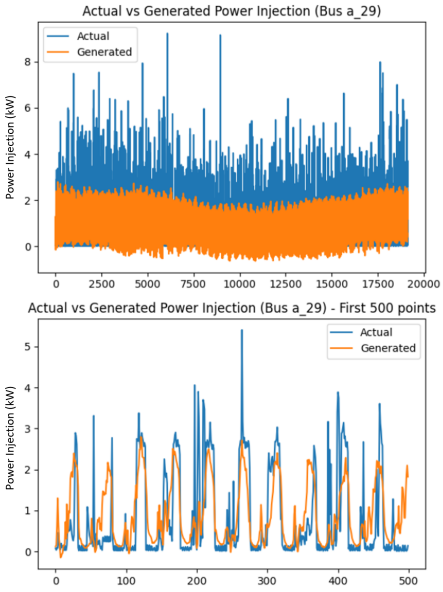
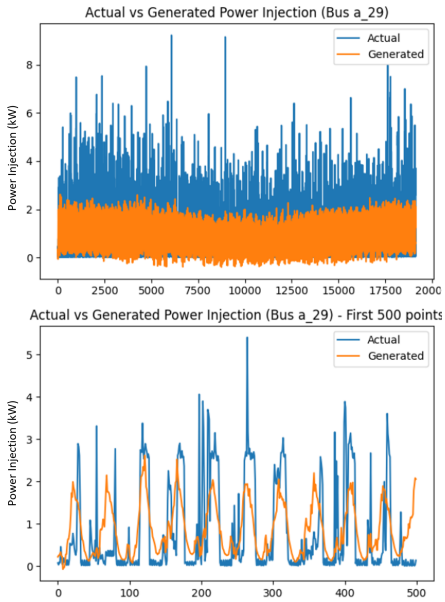
Model	Variables	Graphics	Comments
LR_2	$Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		The inclusion of the injected power at the substation and the substation voltage considerably improves the model's performance throughout the entire dataset. Unlike the previous model, this approach does not lead to significant error propagation, thus maintaining a closer resemblance to the original history.
LR_3	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		This model reaches some peaks that weren't possile to reach before. However, it fails to reach zero because it lacks input variables representing power that can go zero, unlike the previous model, which includes lags that occasionally reach zero.

Table 4.14: Comparison of LR Models for Single-meter Active Power Reconstruction (Part 2)

Three models were also implemented using LSTM's to reconstruct bus 29 Power Injection historical.

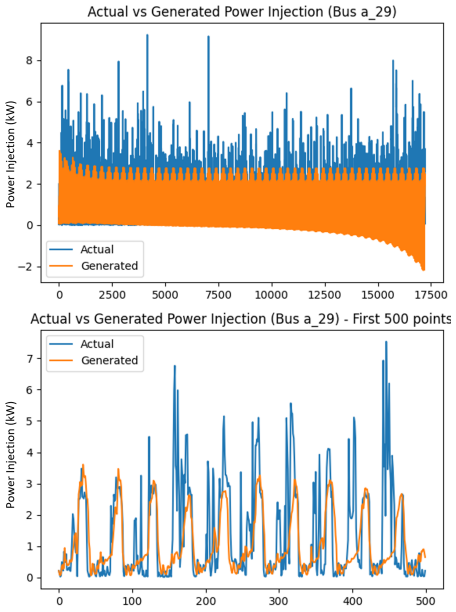
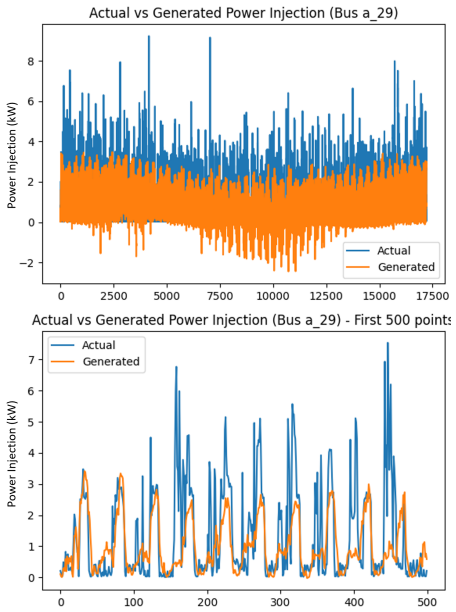
Model	**Variables**	**Graphics**	**Comments**
NN_1	$Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H)		Just like in its equivalent linear regression model, this model can also reconstruct the initial part of the dataset in great detail. However, beyond a certain point, the cyclic variables become dominant, resulting in the emergence of a sinusoidal pattern.
NN_2	$Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		This model effectively captures the behavior of this customer throughout the entire history. However, a common issue with all models is their ability to reach consumption peaks, which are occasional and cannot be explicitly learned by the model.

Table 4.15: Comparison of LSTM Models for Single-meter Active Power Reconstruction (Part 1)

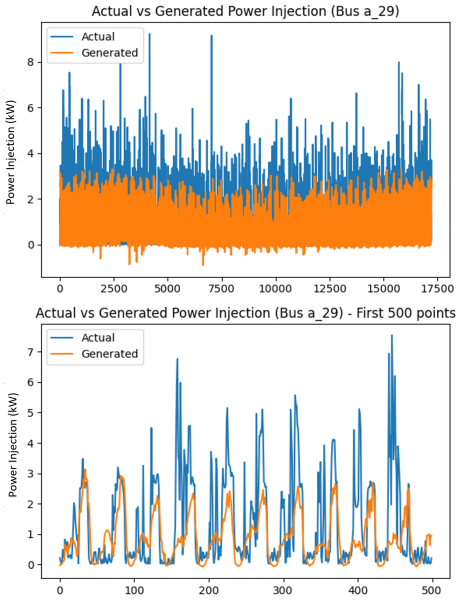
Model	Variables	Graphics	Comments
NN_3	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		Unlike its linear counterpart, this LSTM model without recursive variables can achieve values very close to zero while still maintaining the ability to reach some peaks.

Table 4.16: Comparison of LSTM Models for Single-meter Active Power Reconstruction (Part 2)

To quantitatively assess the most effective model in reconstructing stress history, error metrics were calculated as detailed in Section 4.2. The following results were obtained for LR models and NN models.

Model	Elapsed Time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score	Max Absolute Error
LR_1	293.886	1.114	0.819	0.004	7.603
LR_2	283.962	1.002	0.660	0.161	7.311
LR_3	233.164	0.965	0.734	0.211	7.650
NN_1	1902.419	1.613	1.499	-1.320	7.422
NN_2	2213.369	0.890	0.526	0.233	7.451
NN_3	247.041	0.818	0.626	0.293	7.496

Table 4.17: Error Metrics for Power injection time series reconstruction (single-meter)

NN_2 (Model 2 - LSTM with Real-time Variables and recursive variables) demonstrates the lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), along with the second highest R-squared score among all models. This indicates that NN_2 effectively captures the complex power injection time series patterns, leveraging historical and real-time information. Its competitive max absolute error further solidifies its performance across various aspects of the reconstruction task. The biggest issue regarding this problem is its computation time, which can be longer when opting for other models, namely the NN_3 model with very similar metrics of RMSE, better R2, and is only slightly worse in MAE. However, its computation time can be up to ten times shorter in the case of one bus, which could make it a better solution.

4.5.2 Single-Phase Reconstruction

This section will centre on reconstructing historical data throughout an entire phase, comparing the total phase error across different approaches. Specifically, we will focus on enhanced versions of the top-performing models identified in the previous analysis, LR_2, and LR_3, demonstrating superior performance in reconstructing single-meter data.

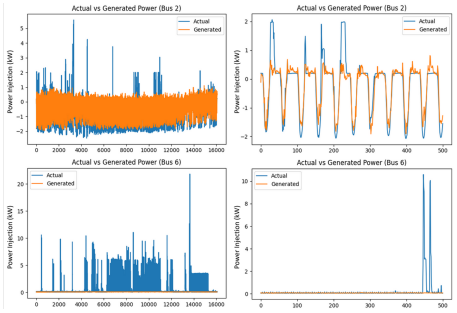
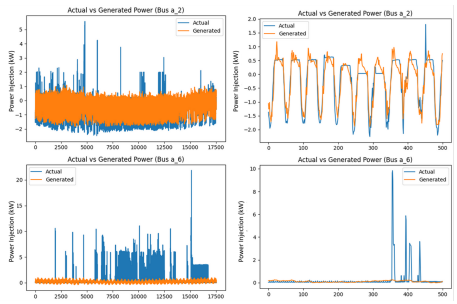
Model	**Variables**	**Graphics**	**Comments**
LR_1	For each bus i in phase a: $Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		This figure illustrate that for reconstructing the historical data of injected active power, this linear model may not be ideal for the presented buses. This is because the apparent lack of daily patterns at some nodes makes simpler models like linear regressions unable to explain their behaviour adequately
LR_2	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		Compared to the previous model, although it shows a better reconstructive line through much of the dataset, it generates significant discrepancies due to the lack of prior information inherent in the recursive model. However, overall, the lower complexity of this model leads to qualitatively better results for a large portion of the dataset.

Table 4.18: Comparison of LR Models for Single-phase Active Power Reconstruction

We will emphasize refined versions of the best-performing LSTM models identified in earlier analyses: NN_2 and NN_3, which exhibited superior performance in reconstructing data from individual meters.

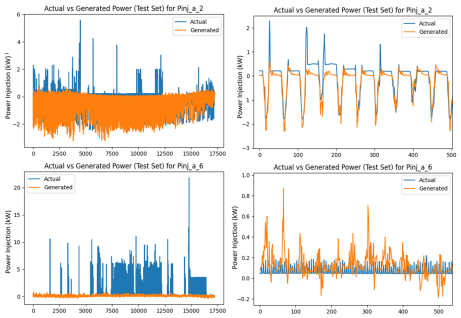
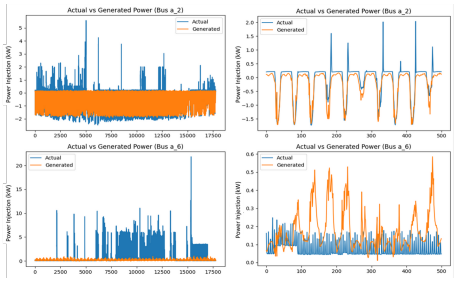
Model	Variables	Graphics	Comments
NN_1	For each bus i in phase a: $Pinj_{29}(t-30m)$ $Pinj_{29}(t-24h)$ $Pinj_{29}(t-48h)$ $Pinj_{29}(t-168h)$ $Pinj_{29}(t-336h)$ Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		Compared to its LR counterpart, this model maintains a balanced estimation throughout the entire dataset, including the final phases where the linear regression model completely diverges. However, a noticeable error remains particularly evident on certain buses, notably bus 6. This is primarily due to bus 6 lacking any discernible pattern, making its accurate reconstruction impossible.
NN_2	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} $Pinj_{a_PT}(t)$		This model can reconstruct some buses with considerable detail, such as Bus 2. However, it fails to accurately reconstruct bus 6, consistently keeping its values around 0.

Table 4.19: Comparison of LSTM Models for Single-phase Active Power Reconstruction

The provided table (Table 4.20) summarizes the performance of four single-phase power injection reconstruction models: LR_1, LR_2, NN_1, and NN_2.

Model	Elapsed Time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score
LR_1	370.132	1.043	0.682	0.038
LR_2	330.553	0.923	0.541	0.221
NN_1	1677.095	0.944	0.561	0.187
NN_2	430.552	0.818	0.469	0.337

Table 4.20: Error Metrics for Power Injection Time Series Reconstruction (single-phase)

Based on the error metrics, it's possible to conclude that NN_2 consistently outperforms the other models in terms of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared Score, making it the best model for single-phase power injection reconstruction in this study. Although NN_1 has the highest computation time, it does not significantly outperform the simpler LR_2 model regarding error metrics. NN_2 strikes a good balance between computation time and accuracy. Between the linear regression models, LR_2 performs better than the recursive LR_1 model, suggesting that adding complexity to the model without addressing inherent data

patterns does not necessarily improve results. Both linear and recurrent models struggle with certain buses that lack clear patterns, such as bus 6. This indicates that regardless of the model complexity, data characteristics play a crucial role in the reconstruction accuracy.

4.5.3 Full Grid Reconstruction

This subsection will focus on presenting the results related to the reconstruction of the Power Injection time series across the entire network. For this purpose, only a LR model was used, as it yielded the best results for reconstructing a single phase (model LR_2). The variables used were:

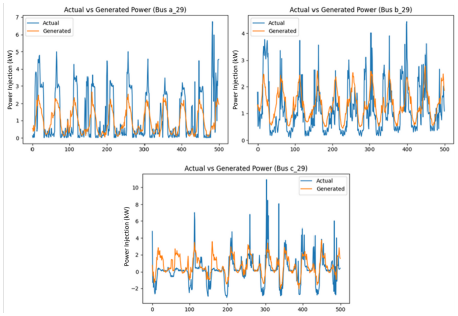
Model	Variables	Graphics	Comments
LR_1	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} V_{b_PT} V_{c_PT} $Pinj_{a_PT}(t)$ $Pinj_{b_PT}(t)$ $Pinj_{c_PT}(t)$		This model present the historical reconstruction results for the same bus across three different phases (bus 29) with Linear Regression reconstruction, allowing us to understand qualitatively the effect of phase differences on historical reconstruction.

Table 4.21: Comparison of LR Models for Full Grid Active Power Reconstruction

Continuing the previous procedure for the historical power injection reconstruction of a single phase, only one model was used for the reconstruction of the entire network history using LSTM's, utilizing the following variables:

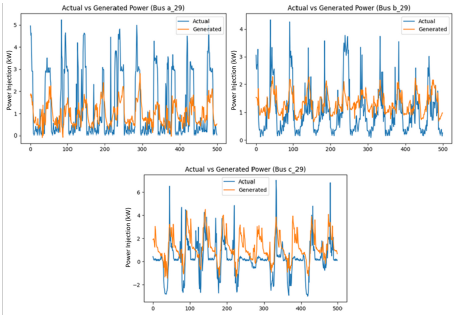
Model	Variables	Graphics	Comments
NN_1	Sin(WD) Cos(WD) Sin(H) Cos(H) V_{a_PT} V_{b_PT} V_{c_PT} $Pinj_{a_PT}(t)$ $Pinj_{b_PT}(t)$ $Pinj_{c_PT}(t)$		This model present the historical reconstruction results for the same bus across three different phases (bus 29) with LSTM's reconstruction, allowing us to understand qualitatively the effect of phase differences on historical reconstruction.

Table 4.22: Comparison of LSTM Models for Full Grid Active Power Reconstruction

The provided table (Table 4.23) summarizes the performance of two Full Grid Power injection reconstruction models: LR_1 and NN_1.

Model	Elapsed time (s)	Root Mean Squared Error	Mean Absolute Error	R-squared Score
LR_1	344.517	0.968	0.646	0.356
NN_1	362.840	0.921	0.580	0.313

Table 4.23: Error Metrics for voltage time series reconstruction (Full Grid)

LR_1 demonstrates superior computational efficiency with a shorter elapsed time of 344.517 seconds compared to NN_1's 362.840 seconds. NN_1 achieves a lower RMSE, MAE only has a slightly lower R2 than LR_1. NN_1's stronger metrics suggest it provides more accurate and reliable power injection reconstructions across the entire dataset.

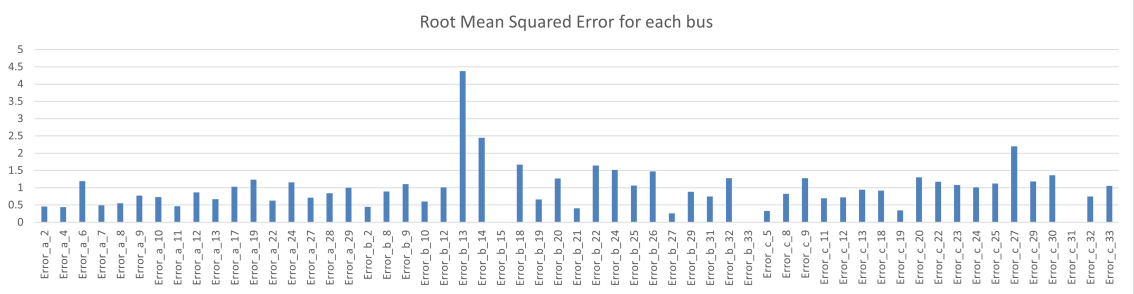


Figure 4.6: Error for each bus using LR for Active Power injection reconstruction

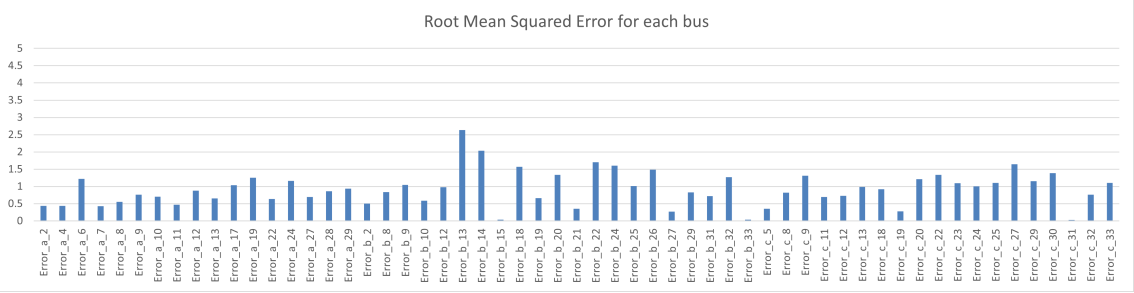


Figure 4.7: Error for each bus using LSTMs for Active Power injection reconstruction

As shown in Figure 4.6 and 4.7, in contrast to the voltage, a decline in reconstruction quality is not inevitable when moving away from the substation. This is because there is no discernible correlation between the power injected from medium to low voltage and the power injected into each bus. Furthermore, there is a notable discrepancy between the reconstruction outcomes obtained through linear regression and those achieved through neural networks. The maximum RMSE value is considerably higher in the linear regression.

4.6 Reconstruction with equipment knowledge

This section describes the improvements made by incorporating equipment information, primarily EV and PV generation data. The model used for the time series reconstruction in this analysis

was the one with the best metrics for a single meter, specifically model NN_3 from Single-phase power injection reconstruction. The initial analysis was conducted for bus a_29 and considered the addition of information regarding the consumption from EV chargers.

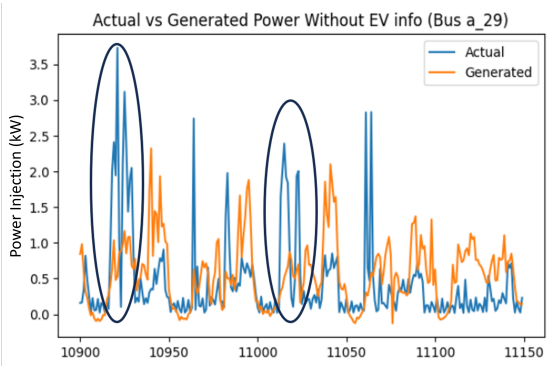


Figure 4.8: Power Injection Baseline Reconstruction

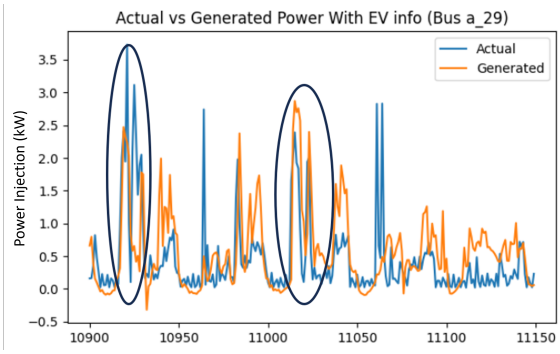


Figure 4.9: Power Injection Reconstruction with EV Charger Consumption Information

As can be seen from Figure 4.8 to Figure 4.9, the model provides a more straightforward explanation for the two surrounded peaks, which precisely corresponds to two consecutive days of charging by this customer.

Model	Elapsed time (s)	Root mean squared error	Mean absolute error	R-squared score	Max Absolute error
Baseline	163.936	0.914	0.613	0.279	7.341
Baseline + EV info	154.41	0.861	0.569	0.36	7.176

Table 4.24: Error Metrics for Power Injection Time Series Reconstruction with and without EV charging power

Incorporating EV charger consumption data demonstrably improves the power injection time series reconstruction for bus *a*_29. This translates to a clearer understanding of power consumption patterns, potentially leading to better grid management and planning for future infrastructure needs. All error metrics (Root Mean Squared Error, Mean Absolute Error, Max Absolute Error) show a decrease with EV information, signifying a more accurate fit.

The second analysis was conducted for bus a_28 to understand the effect of adding information about PV generation for better reconstruction compared to the base reconstruction.

Two approaches were considered for handling the information about photovoltaic generation. First, the initial model was augmented with the irradiance for the specific location to enlighten the model on its capacity to inject power due to the presence or absence of panels. The second approach involved reconstructing the historical photovoltaic generation considering the same database divisions as for the main problem, using the following variables:

$$\text{PV Model:} \begin{cases} \text{Sin(WD)} \\ \text{Cos(WD)} \\ \text{Sin(H)} \\ \text{Cos(H)} \\ \text{Irradiance} \end{cases}$$

The results of this PV Generation Reconstruction are as follows:

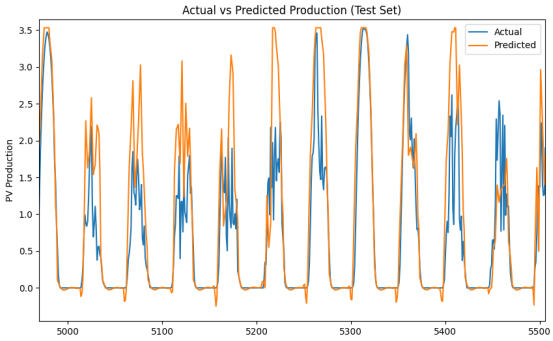


Figure 4.10: PV Generation Reconstrunction

The quantitative metrics of this method are presented in table 4.25.

Metric	**Value**
Mean Squared Error (MSE)	0.277
Mean Absolute Error (MAE)	0.306
R-squared (R2) Score	0.758
Max Absolute Error	3.059

Table 4.25: Error Metrics for Photovoltaic Generation Reconstruction

The following two ways of incorporating information into the model were compared, one using only irradiance and the other using the generated PV history.

Model	Elapsed time (s)	Root mean squared error	Mean absolute error	R-squared score	Max Absolute error
Baseline Reconstruction	223.251	0.846	0.570	0.661	6.119
Baseline Reconstruction + Irradiance	220.715	0.838	0.561	0.667	6.022
Baseline Reconstruction + PV Model	354.348	0.830	0.533	0.674	6.21

Table 4.26: Error Metrics for Power Injection Time Series Reconstruction

The *Baseline Reconstruction + PV Model* appears to be the best-performing method based on RMSE, MAE, and R-squared score. However, it also has the longest computation time (354.348 seconds). The *Baseline Reconstruction + Irradiance* balances accuracy and training time well. The inclusion of PV information in the model will increase the accuracy of the Power Injection Reconstruction model.

The third and final analysis of this subsection focuses on comparing the inclusion of the two types of information, PV and EV, separately and together in the baseline model and verifying the results obtained for each of them; this analysis was conducted to bus *c*_29.

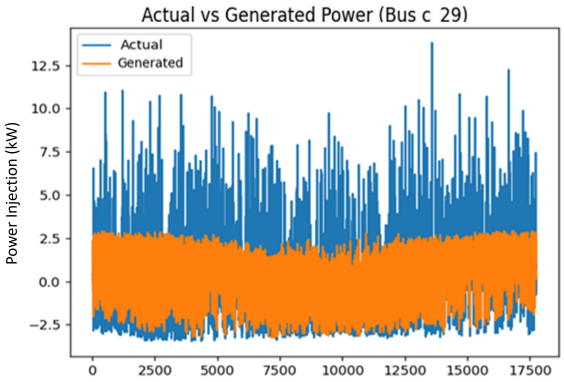


Figure 4.11: Baseline Reconstruction bus *c*_29

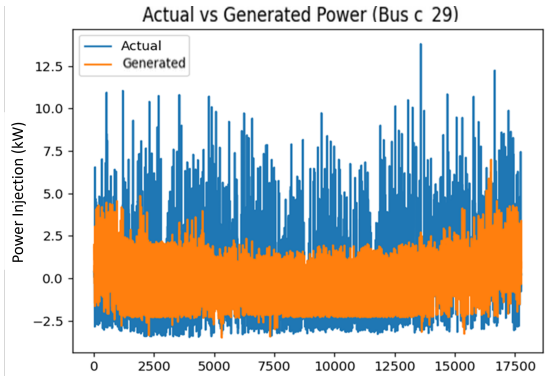


Figure 4.12: Baseline Reconstruction with EV information bus *c*_29

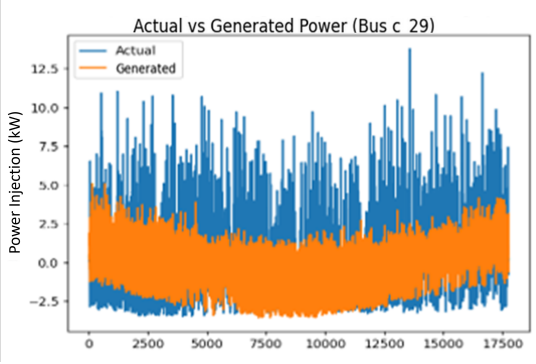


Figure 4.13: Baseline Reconstruction with PV information bus *c*_29

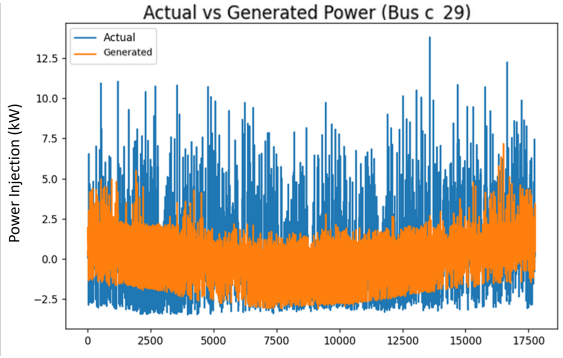


Figure 4.14: Baseline Reconstruction with EV and PV information bus *c*_29

It can be observed that from Figure 4.11 to Figure 4.12, some power peaks are better reconstructed. This result is precisely due to the incorporation of EV information into the model, which leads to an overall improvement in power injection reconstruction.

It is also possible to conclude that from Figure 4.11 to Figure 4.13, the lower peaks corresponding to photovoltaic production will be better reconstructed, particularly in the central zone of the historical data corresponding to summer months, when photovoltaic production is higher.

In Figure 4.14, both pieces of information can be integrated, significantly improving the baseline reconstruction.

Model	Elapsed time (s)	Root mean squared error	Mean absolute error	R-squared score	Max Absolute error
Baseline Reconstruction	151.241	1.195	0.759	0.475	12.284
Baseline + PV info	331.687	1.134	0.658	0.528	11.396
Baseline + EV info	176.542	1.132	0.685	0.529	11.091
Baseline + PV info + EV info	324.348	1.127	0.658	0.534	11.128

Table 4.27: Error Metrics for Power Injection Time Series Reconstruction

Based on the error metrics provided in Table 4.27, including both PV and EV information (*Baseline + PV info + EV* info) in the *Baseline* model leads to the best overall performance. This is evident from the lowest root mean squared error (RMSE) of 1.127, the lowest mean absolute error (MAE) of 0.658, and the highest R-squared score of 0.534. In conclusion, leveraging both

PV and EV information offers the most accurate reconstruction of the power injection time series, as demonstrated by the error metrics and the visual comparison of the figures.

4.7 Voltage reconstruction considering statistical properties

In this section, a detailed examination of the statistical properties of the actual data time series and the time series generated by the proposed model is carried out. The comparative analysis aims to verify how well the model manages to capture and reproduce the essential statistical characteristics of the original time series. For this analysis, two buses, a_7 and a_29, were considered. The data used in this section was generated using the best single-meter model for the voltage analysis, which was considered model LR_3 for single-meter reconstruction.

Bus a_7 Voltage

The following figures show the voltage generated data and the actual data of bus a_7.

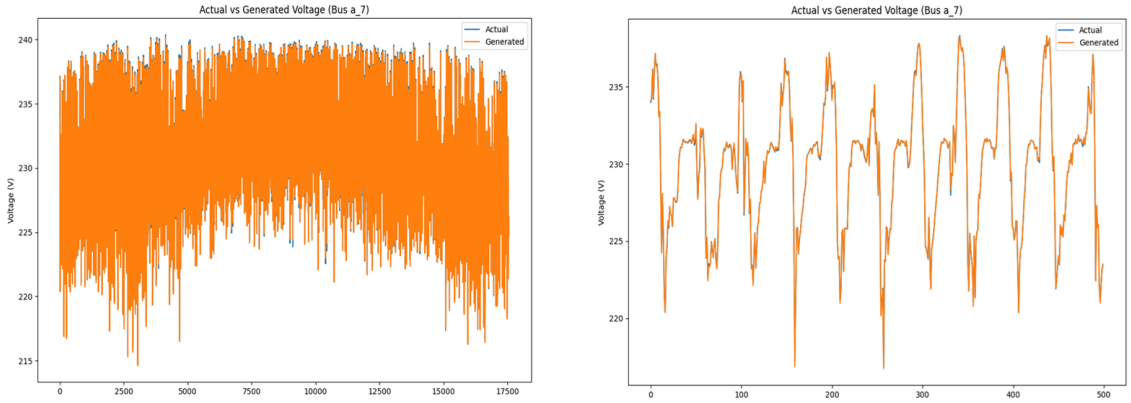


Figure 4.15: Voltage Reconstruction Bus a_7

The model used to reconstruct the voltage profiles works very well, as seen from the near overlap between the generated and real series. A more quantitative analysis of the comparison of the series reconstructions is given in section 4.4.

The first set of characteristics considered were the mean, standard deviation, skewness, kurtosis, and correlation.

Property	Real Values	Generated Values
Mean	231.438	231.432
Standard Deviation	3.975	3.958
Skewness	-0.091	-0.134
Kurtosis	0.026	0.070
Correlation	0.999	

The analysis of statistical properties such as mean, standard deviation, skewness, kurtosis, and correlation demonstrates that the model used to reconstruct the voltage profiles of bus a_7 performs exceptionally well.

Firstly, the mean values of the generated and actual data are nearly identical, indicating that the central tendencies of both datasets are well-aligned. This shows that the model accurately captures the average voltage levels of bus a_7.

Secondly, although the standard deviations of the generated data are slightly lower than those of the actual data, the difference is minimal. This similarity suggests that the model effectively replicates the variability in voltage measurements, maintaining the spread of data points around the mean.

Thirdly, there are minor differences in skewness and kurtosis between the generated and actual data, but both distributions display negative skewness and positive kurtosis. This indicates that they have similar shapes, with moderate asymmetry and peakedness. This consistency suggests that the model preserves the key characteristics of the voltage distribution.

Lastly, the high correlation coefficient (0.9997) between the generated and actual values indicates a strong linear relationship. This high correlation suggests that the model closely follows the overall trend and pattern of the actual voltage data of bus a_7.

The small differences observed in statistical properties between the generated and actual series suggest that the model accurately reconstructs the voltage profiles of bus a_7. These findings highlight the model's reliability and effectiveness in capturing key features of the voltage data, validating its practical utility for accurate voltage reconstruction.

A further method of elucidating the relationship between the actual and generated values is a detailed analysis of the relevant PDFs.

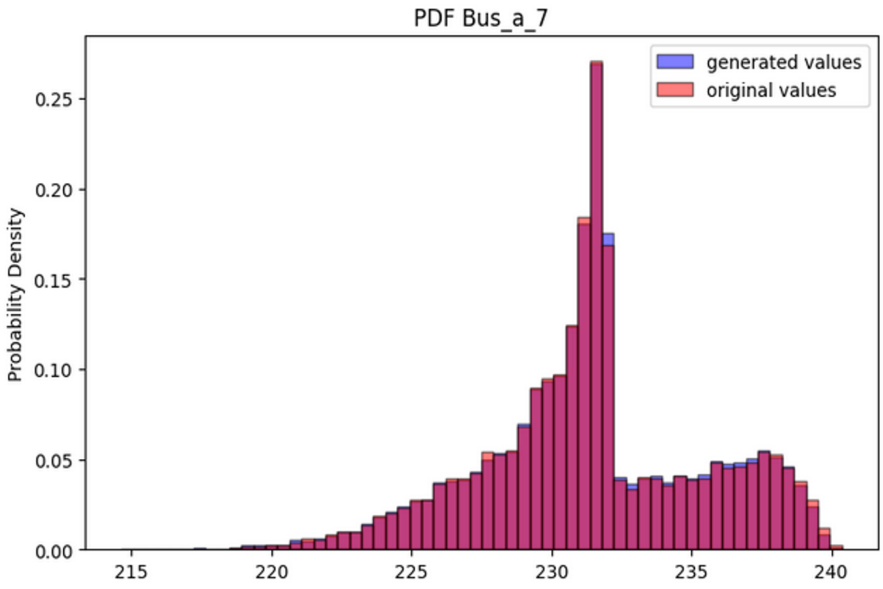


Figure 4.16: Probability Density Function comparison Bus a_7

The provided probability density function (PDF) plot compares the generated and original voltage values distribution for bus a_7, revealing several key conclusions about the model's performance. The PDF plot indicates that the generated values closely align with the original values across the entire range, suggesting that the model accurately reproduces the overall distribution of the original voltage data. The generated and original data exhibit a prominent peak around the 230-235 range, with the highest density occurring just below 235. This alignment of the central peak demonstrates that the model effectively captures the central tendency of the voltage distribution.

Additionally, the distribution's tails at the lower end (around 220-225) and the higher end (235-240) show a good match between the generated and original values. This similarity indicates that the model successfully captures the spread and extremities of the voltage data, maintaining accuracy across the entire spectrum. While there are minor differences in the heights of some bars, particularly at the central peak and around certain mid-range values, these deviations are minimal and do not significantly affect the overall alignment. This suggests that the model maintains a high level of accuracy in replicating the original data's distribution.

Furthermore, the probability density of both datasets is very close, with slight variations that fall within an acceptable range. This consistency in probability density further validates the model's effectiveness in replicating the statistical properties of the original data. The PDF plot demonstrates that the model accurately reproduces the original voltage data distribution for bus a_7. The generated values closely align with the original values across the entire range, with only minor differences that do not significantly impact the match. This high level of agreement underscores the model's reliability and effectiveness in capturing the voltage data's key statistical characteristics, validating its practical utility for accurate voltage reconstruction.

The following analysis was carried out to analyze the autocorrelation function (ACF) and the partial autocorrelation function (PACF).

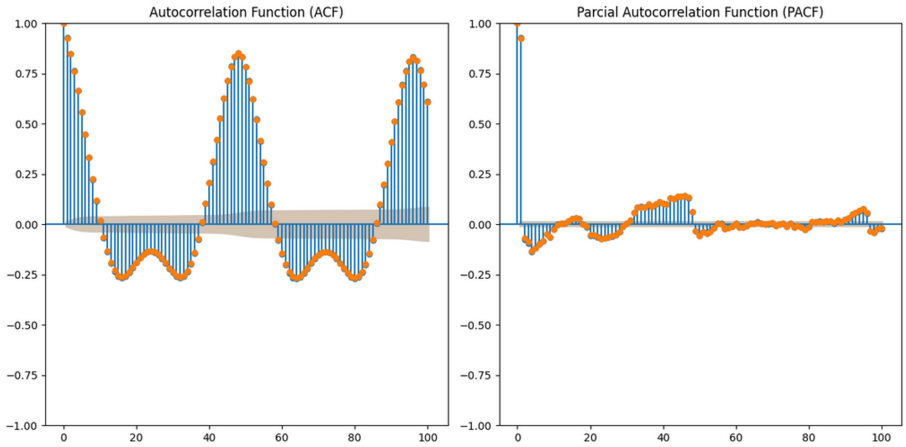


Figure 4.17: ACF and PACF comparison Bus a_7

The provided plots show the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) for the original (blue) and generated (orange) voltage values for bus a_7, allowing us to draw several conclusions about the model's performance.

The ACF plot illustrates the correlation of the time series with its own lagged values. The original data (blue) exhibits a clear periodic pattern, with significant peaks at regular intervals, indicating strong seasonality or periodicity in the data. The generated data (orange) closely follows this pattern, with similar peaks at the same intervals. The alignment between the original and generated ACF suggests that the model successfully captures the original data's temporal dependencies and periodic structure.

After removing the effects of earlier lags, the PACF plot shows the correlation of the time series with its own lagged values. The original data (blue) shows a significant spike at lag 1 and then fluctuates around zero, indicating that the first lag can explain most of the autocorrelation. The generated data (orange) also shows a significant spike at lag 1 and follows a similar pattern to the original data for subsequent lags. This indicates that the model effectively captures the primary autocorrelation structure of the original data.

The close alignment of the ACF and PACF plots for the original and generated data suggests that the model accurately replicates the autocorrelation and partial autocorrelation structures of the original voltage data for bus a_7. The ability of the model to capture both the overall periodicity and the primary autocorrelation characteristics validates its effectiveness in preserving the temporal dependencies inherent in the original data. This high level of agreement indicates that the generated data is a reliable representation of the original series, suitable for further analysis and practical applications.

The next and final statistical analysis in this subsection analyzes the trend, seasonality, and residuals of the generated series compared to the real series.

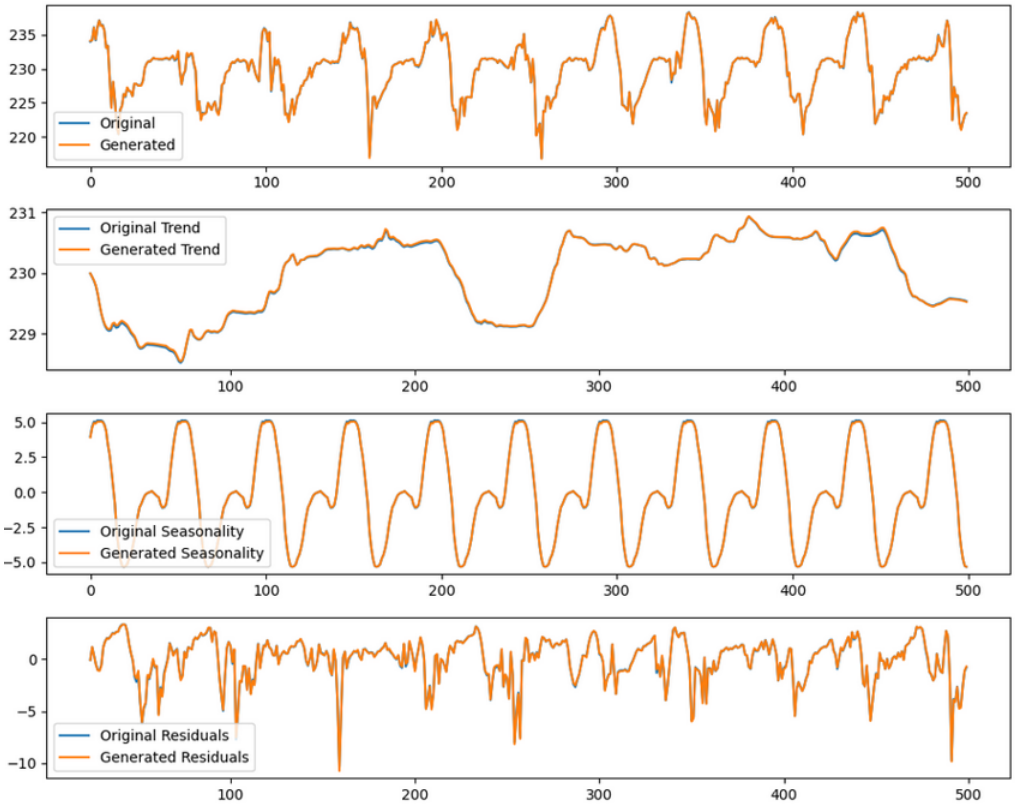


Figure 4.18: Serie decomposition Bus a_7

The top plot of Figure 4.18 compares the original time series with the generated time series. Both series closely follow each other, indicating that the generated series matches the original series well.

The second plot of Figure 4.18 compares the trend component of the original time series with that of the generated time series. The trends in both series align well, demonstrating that the generated data captures the long-term movement and patterns of the original data.

The third plot of Figure 4.18 compares the seasonal component of the original time series with that of the generated time series. The seasonal patterns are almost identical, indicating that the generated series accurately replicates the periodic fluctuations of the original series.

The bottom plot of Figure 4.18 compares the residuals (the noise component after removing the trend and seasonality) of the original time series with that of the generated time series. The residuals also show a high degree of similarity, suggesting that the random variations or noise in the generated series are very similar to those in the original series.

The generated time series data closely replicates the original time series regarding overall patterns, trends, seasonal components, and residuals. This indicates that the generation method used to reconstruct data is highly effective at mimicking the characteristics of the original time series.

Bus a_29 Voltage

The following figures show the voltage generated data and the actual data of bus a_29.

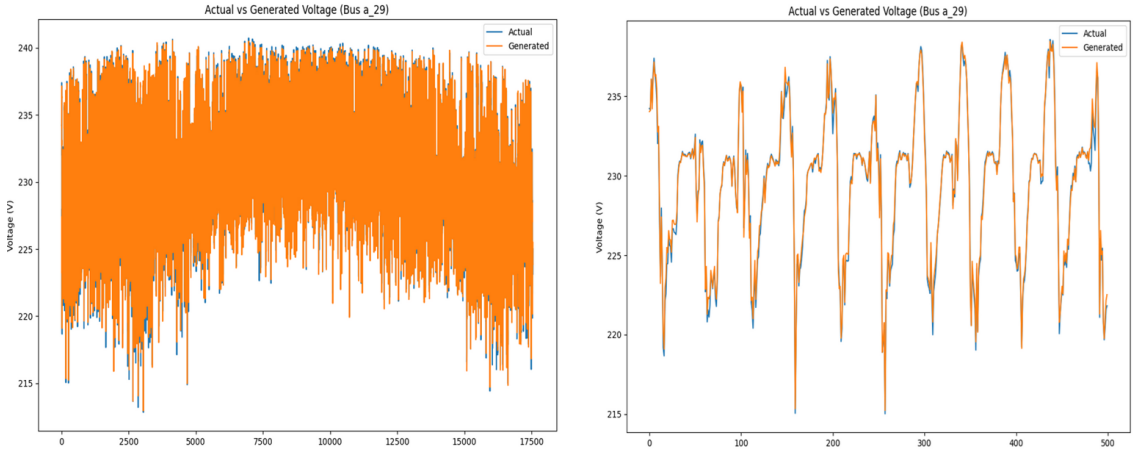


Figure 4.19: Voltage Reconstruction Bus *a*_29

The model used to reconstruct the voltage profiles works very well, as seen from the near overlap between the generated and real series. A more quantitative analysis of the comparison of the series reconstructions is given in section 4.4.

The first set of characteristics considered were the mean, standard deviation, skewness, kurtosis, and correlation.

Property	**Real Values**	**Generated Values**
Mean	231.136	231.141
Standard Deviation	4.308	4.260
Skewness	-0.217	-0.178
Kurtosis	0.159	0.096
Correlation	0.996	

The analysis of statistical properties, including mean, standard deviation, skewness, kurtosis, and correlation, reveals that the model used for reconstructing the voltage profiles of bus a_29 performs exceptionally well.

Firstly, the mean values of the generated and actual data are very close, indicating that the central tendencies of both datasets align closely. This suggests that the model accurately captures the average voltage levels of bus a_29.

Secondly, the standard deviations of the generated data are slightly lower than those of the actual data, but the difference is minimal. This similarity implies that the model effectively replicates the variability in voltage measurements, preserving the spread of data points around the mean.

Thirdly, while slight differences in skewness and kurtosis exist between the generated and actual data, both distributions exhibit negative skewness and positive kurtosis, indicating comparable shapes with moderate asymmetry and peakedness. This similarity suggests that the model maintains the essential characteristics of the voltage distribution.

Lastly, the high correlation coefficient between the generated and real values (0.9964) indicates a strong linear relationship. This high correlation suggests that the model closely follows the overall trend and pattern of the actual voltage data of bus a_29.

In summary, the small differences observed in statistical properties between the generated and actual values suggest that the model accurately reconstructs the voltage profiles of bus a_29. These findings underscore the reliability and effectiveness of the model in capturing key features of the voltage data, validating its utility for practical applications requiring accurate voltage reconstruction.

A further method of elucidating the relationship between the actual and generated values is a detailed analysis of the relevant PDFs.

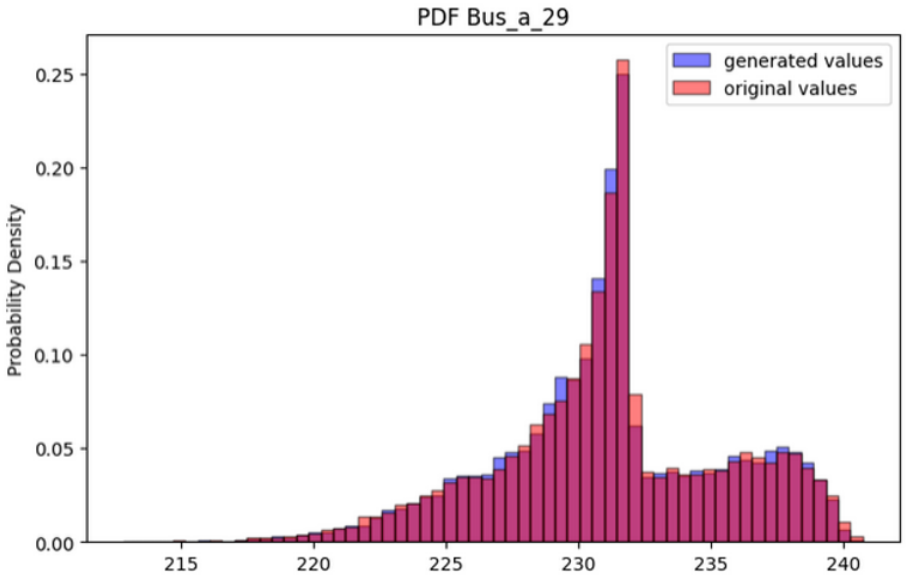


Figure 4.20: Probability Density Function comparison Bus *a*_29

A comparison of the probability density functions (PDFs) of the original and generated data for Bus a_29 reveals that both distributions exhibit a very similar overall shape. This indicates that the generated data accurately captures the general trend of the original data. The peaks of both distributions are centered around the same range, particularly close to 230, as it was expected, suggesting that the generated values' central tendency (mean) aligns well with that of the original values. The width of the distributions appears to be very similar, indicating that the variability (spread) of the generated values is comparable to that of the original values. This suggests that the range of fluctuations observed in the original data set is reflected in the generated data set. The tails of both series are also very similar on both sides.

The following analysis was carried out to analyze the autocorrelation function (ACF) and the partial autocorrelation function (PACF).

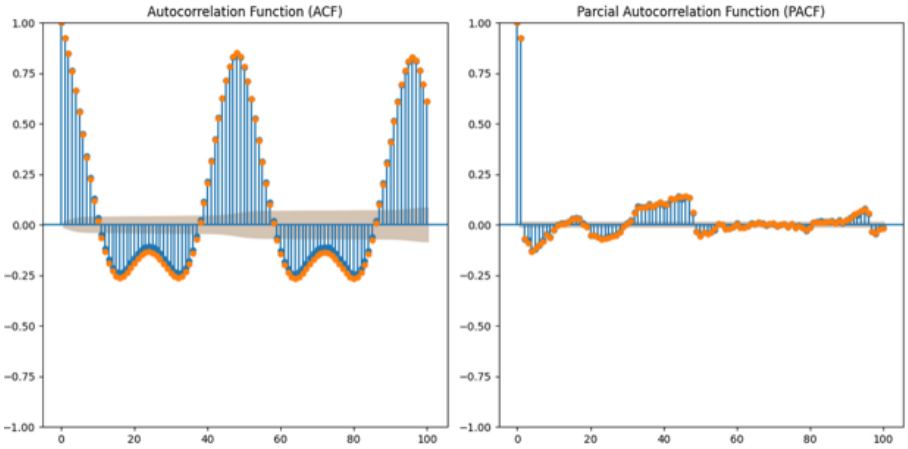


Figure 4.21: ACF and PACF comparison Bus a_29

The ACF plot for the original series shows a very clear and strong oscillatory pattern. This indicates the presence of a significant seasonal component with a well-defined periodicity. The amplitude of these oscillations remains high throughout the entire graph, highlighting a persistent and regular pattern in the data.

The generated series, depicted in orange, also displays an oscillatory pattern closely mimics the original series. This suggests that the generated series has successfully captured the seasonality of the original series. The amplitude of the oscillations in the generated series is similar to that of the original series, although there are slight variations at certain points. These variations could indicate minor differences in the replication of the periodicity.

When examining the PACF plot for the original series, we observe that the first lag has a high value, followed by smaller values that hover around zero. There are significant partial autocorrelations at the initial lags, but these dissipate quickly. This suggests that the original series has strong dependencies at the initial lags, which diminish rapidly.

In contrast, the PACF plot for the generated series also shows a high value at the first lag, similar to the original series. However, the generated series displays significant partial autocorrelations at the initial lags, with some subsequent fluctuations that are not as prominent in the original series. Additionally, there appear to be more lags with significant partial autocorrelations compared to the original series. This indicates potential differences in the dependency structure between the two series.

The next and final statistical analysis in this subsection analyzes the trend, seasonality, and residuals of the generated series compared to the real series.

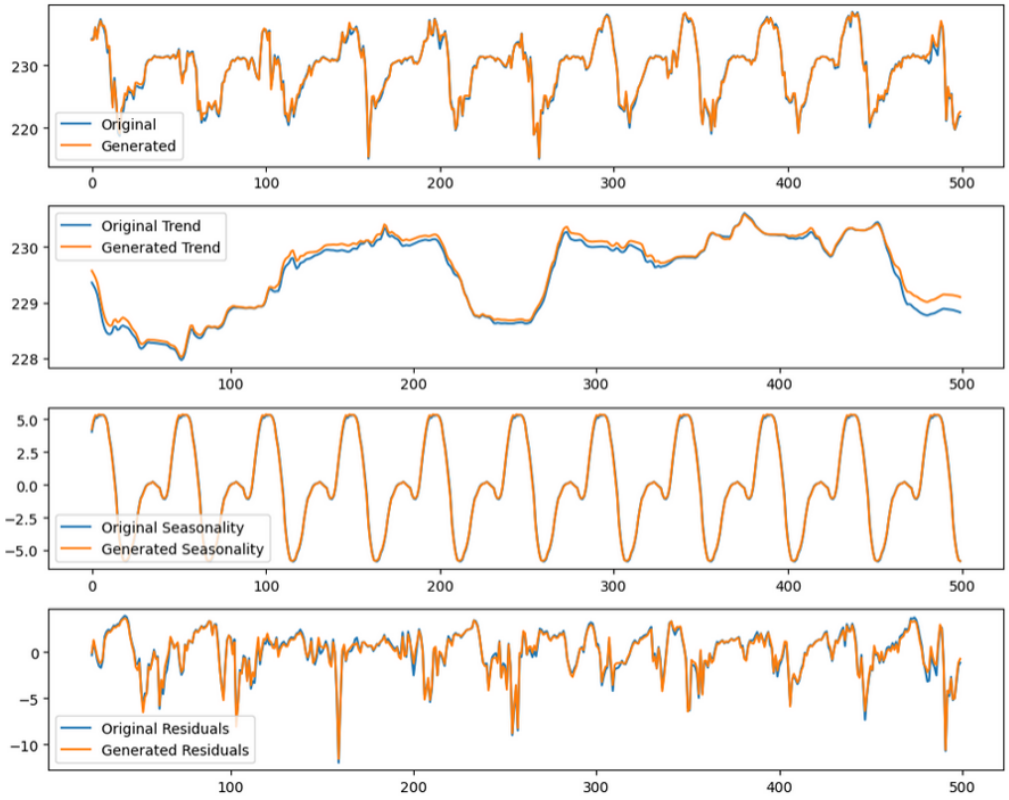


Figure 4.22: Serie decomposition Bus a_29

The top plot of Figure 4.22 shows a remarkable alignment between the original and generated time series data, indicating that the model successfully captures the overall pattern and behaviour of the original series. The close match implies that the generated data accurately represents the key characteristics of the original time series.

The second plot of Figure 4.22 compares the trends of the original and generated data, revealing a strong similarity with only minor deviations. This suggests that the model effectively captures the underlying trend, though some fine-tuning might be necessary to eliminate slight discrepancies around certain peaks and troughs.

The third plot of Figure 4.22 illustrates the seasonal components of the original and generated data, showing an almost perfect alignment. This indicates that the model has accurately replicated the periodic fluctuations of the time series, demonstrating its capability to understand and reproduce the cyclic nature of the data.

The bottom plot of Figure 4.22 compares the original and generated data residuals, with a close match observed. This suggests that the model effectively captures the random noise and variability present in the original time series, ensuring that the generated data not only mimics the overall patterns but also the inherent randomness of the original data.

This level of accuracy in reconstructing time series data can be highly valuable for scenarios where additional data is needed for analysis, testing, or simulation, and using the original data is not feasible due to privacy or availability constraints.

4.8 Power reconstruction considering statistical properties

This section presents a comprehensive analysis of the statistical properties of the actual data time series, the time series generated with the baseline power reconstruction model with EV and PV information, and the statistically preserved power reconstruction (SPPR) model. The objective of the comparative analysis is to ascertain the extent to which the model can accurately capture and reproduce the fundamental statistical properties of the original time series. For the purposes of this analysis, two buses, a_7 and a_29, were selected for investigation. The data employed in this section was generated using the optimal single-meter model for power injection analysis, which was identified as model NN_3 for single-meter reconstruction in the *Baseline* model and the same model with the loss function modification for the subsequent model.

Bus a_7 Power

The following figures show the power injection generated data and the actual data of bus a_7.

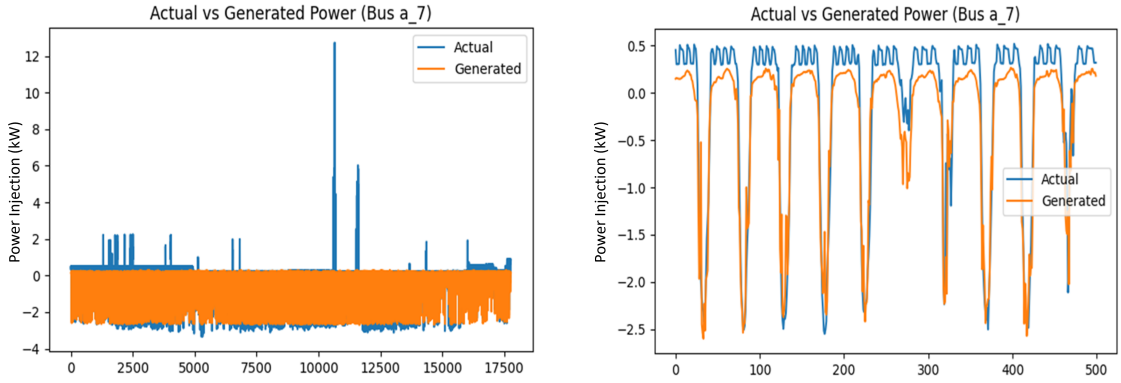


Figure 4.23: Power Injection Baseline Reconstruction Bus *a_7*

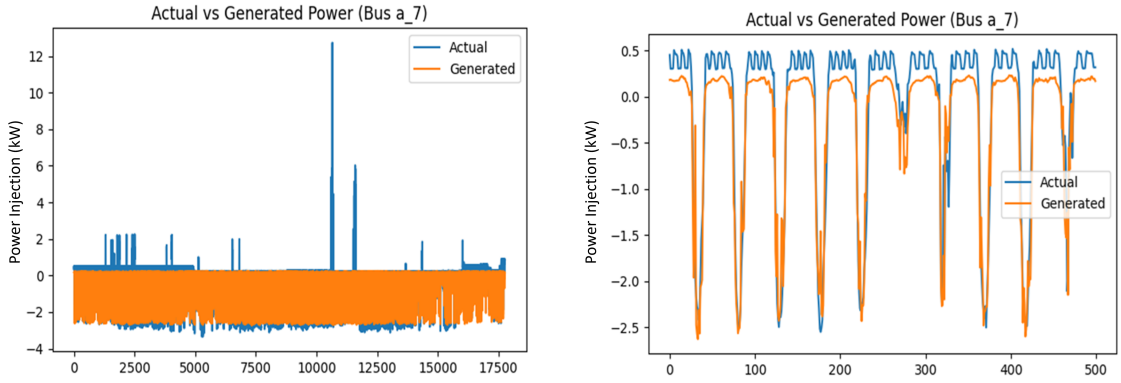


Figure 4.24: SPPR model Bus *a_7*

Both models work particularly well for buses with high photovoltaic production. This is because photovoltaic production has very well-defined daily patterns.

The first set of characteristics considered were the mean, standard deviation, skewness, kurtosis, and correlation.

Property	**Real Values**	**Generated Values Baseline model**	**Generated Values SPPR model**
Mean	-0.432	-0.540	-0.522
Standard Deviation	1.080	0.989	1.001
Skewness	-0.645	-1.061	-1.095
Kurtosis	2.378	-0.525	-0.497
Correlation	-	0.93	0.93

Table 4.28: Comparison of statistical properties between real and generated values from different models.

Several observations can be made based on comparing the mean, standard deviation, skewness, and kurtosis between the real and generated values.

Both generated values means are slightly lower than the real values, with the *SPPR* model producing a mean closer to the real values than the *Baseline* model.

The *SPPR* model resembles the real data's variability more closely than the *Baseline* model.

Both generated datasets exhibit more negative skewness than the real values, with the *Baseline* model showing a skewness closer to the real data.

The *SPPR* model kurtosis of -0.49703 is closer to the real values but still represents a distribution with lighter tails.

The correlation between the real values and both sets of generated values is high at 0.93, indicating a strong linear relationship. Both models capture the linear trend well, but differences in mean, standard deviation, skewness, and kurtosis reflect areas where each model deviates from the statistical characteristics of the real data.

In summary, while both generated data models effectively capture the linear relationship with the real data, the *SPPR* model matches the real data's statistical characteristics more closely than the *Baseline* model. The *SPPR* model achieves a mean, Standard Deviation, and kurtosis that are more in line with the real data, though differences still exist in distribution shape and variability.

A further method of elucidating the relationship between the actual and generated values is a detailed analysis of the relevant PDFs.

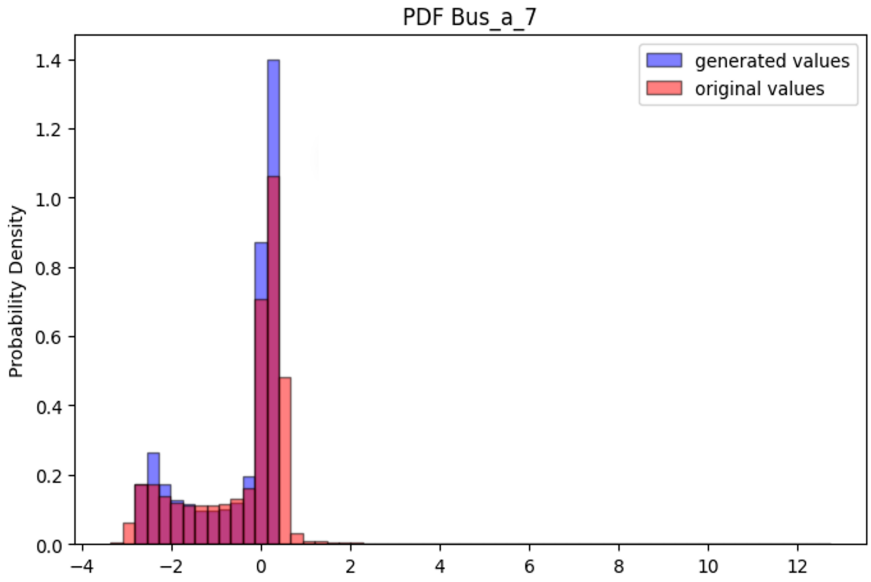


Figure 4.25: Baseline model Probability Density Function comparison Bus *a*_7

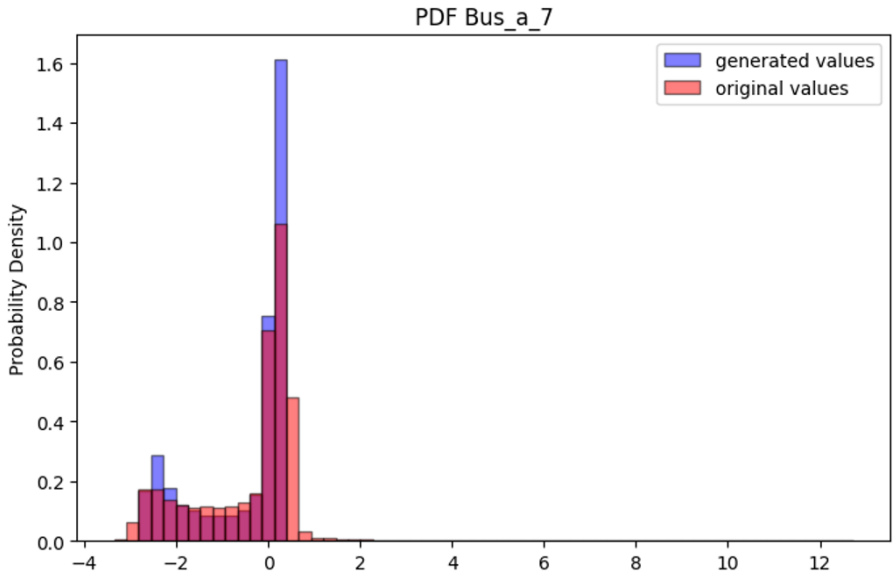


Figure 4.26: SPPR model Probability Density Function comparison Bus *a*_7

A comparison of the probability density functions (PDFs) of the original and both generated datasets for Bus a_7 reveals that the three distributions exhibit a similar overall shape. This indicates that the generated data accurately captures the general trend of the original data. The peaks of the distributions are centred around the same range, particularly close to zero, suggesting that the generated values' central tendency (mean) aligns well with that of the original values. The width of the distributions appears to be similar, indicating that the variability (spread) of both generated values is comparable to that of the original values. This suggests that the range of fluctuations

observed in the original data set is reflected in the generated data set. Both distributions demonstrate a precipitous decline following the peak. The data generated using the *Baseline* model have a smaller peak in the mean value compared to the *SPPR* model precision since it tries to reduce the error in the 0.5 quantile that leads to this effect. Still, it is possible to see that in the previous bar, this model manages to match the real ones much better than the baseline model.

The following analysis was conducted to analyze the autocorrelation function (ACF) and the partial autocorrelation function (PACF).

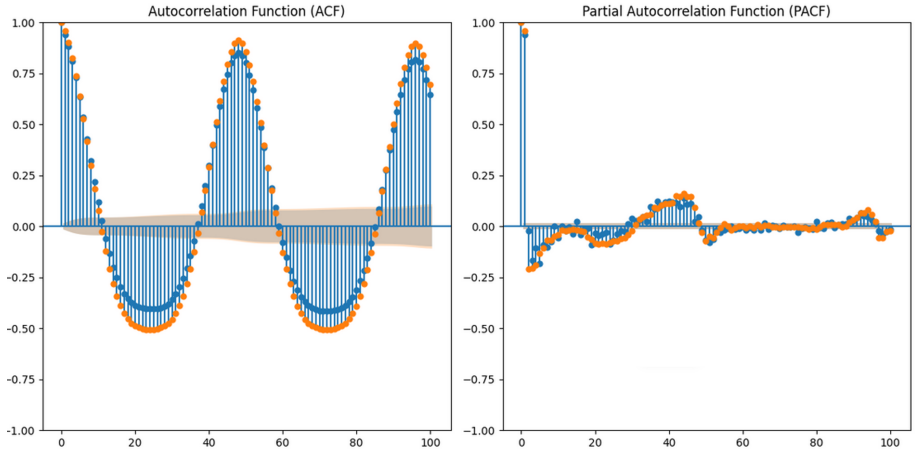


Figure 4.27: Baseline model ACF and PACF comparison Bus *a*_7

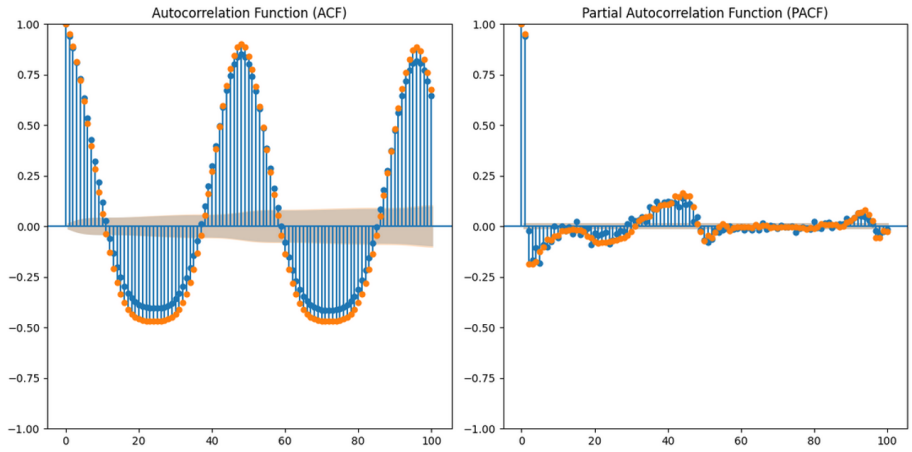


Figure 4.28: SPPR model ACF and PACF comparison Bus *a*_7

The ACF plot shows the correlation of the time series with its own lagged values. The plot exhibits a clear sinusoidal pattern, indicating a strong periodicity in the data. The peaks occur approximately every 48 lags, suggesting a seasonality period of around 48-time units (the measurements are 30 minutes spaced, so 48-time units equals 1 day). The values are quite significant and oscillate between positive and negative, characteristic of a seasonal pattern. The confidence intervals are shown as a shaded area, and many of the lags have autocorrelation values exceeding these bounds, further confirming the presence of significant autocorrelation.

The PACF plot shows the correlation of the time series with its own lagged values, with the effect of shorter lags removed. Unlike the ACF, the PACF shows significant values only at specific lags. The first few lags (up to around lag 10) have significant values, indicating that these lags have a strong partial autocorrelation with the time series. Beyond these initial lags, the values generally stay within the confidence intervals, suggesting that the seasonal effect seen in the ACF is mainly due to the periodicity rather than a direct partial autocorrelation.

Analysis reveals that the autocorrelation function (ACF) plot for values generated by the *SPPR* model more closely matches the actual values, particularly in the lower peaks, compared to those produced by the baseline model. This suggests that the *SPPR* model more accurately captures the temporal dependencies and lag relationships present in the data.

The next and final statistical analysis in this subsection is to analyze the trend, seasonality, and residuals of the generated series compared to the real series.

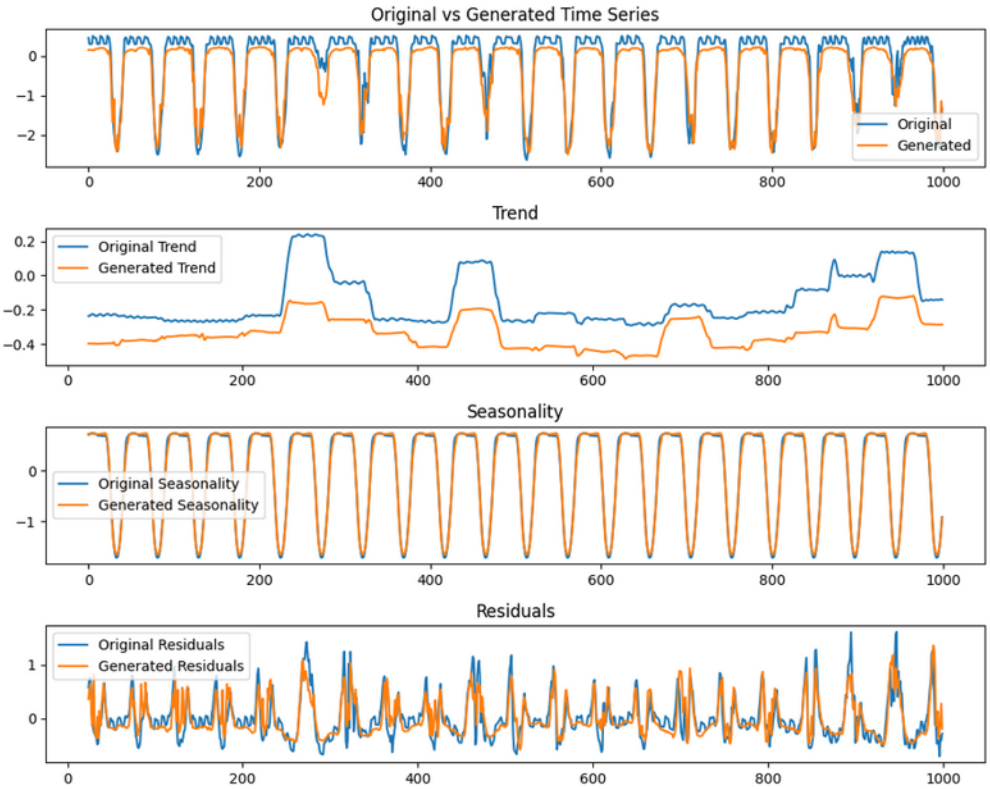


Figure 4.29: Baseline Reconstrunciton Serie decomposition Bus *a*_7

The top plot of Figure 4.29 shows the original time series compared to the generated ones. The generated series captures the general pattern of the original series, showcasing an ability to reflect the overall periodic behaviour. While there are some discrepancies, the alignment of general trends indicates that the model is on the right track and has successfully identified the underlying seasonal structure of the data.

The second plot of Figure 4.29 compares the trends of the original and generated time series.

Although the generated trend is smoother, it successfully captures the overall direction of the original trend. This smoothness can be advantageous in situations where a generalized trend is desired. By effectively highlighting the broad movement of the data, the generated trend demonstrates the model's solid understanding of the trend component.

The third plot of Figure 4.29 illustrates the seasonality component for both the original and generated time series. The generated seasonality closely follows the repeating pattern of the original series, accurately capturing its periodic nature. This indicates that the model has a strong grasp of the seasonal aspect of the data, reflecting its ability to replicate cyclical patterns effectively.

The bottom plot of Figure 4.29 shows the residuals, representing the differences between the observed and generated values. The generated residuals exhibit a similar variability to the original residuals, suggesting that the model captures a substantial portion of the underlying variability. This alignment in variability patterns indicates that the model effectively accounts for a significant amount of the noise in the data.

The generated data approximates the real data's distribution fairly well, capturing the general shape, central tendency, and spread. However, it slightly underrepresents the extreme values and the kurtosis of the real data. This suggests that while the generated data is a good representation, further refinement might be necessary to improve its ability to capture the full extent of the original data's variability, particularly in the tails and peak sharpness.

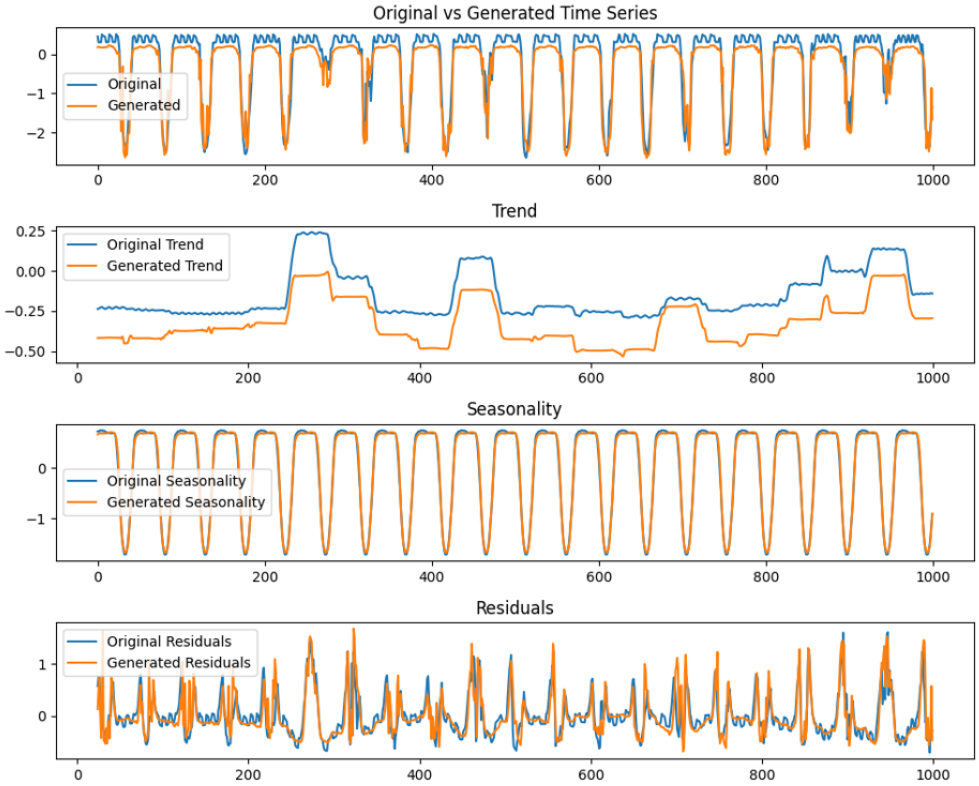


Figure 4.30: SPPR Serie decomposition Bus a_7

The decomposition of the time series obtained using both the baseline and the *SPPR* models allows us to conclude that the *SPPR* model better preserves the original trend and results in residuals closer to the original residuals.

We can infer that for this specific bus, due to its high photovoltaic production, which makes capturing its injected power patterns less challenging, the *SPPR* approach yields superior results compared to the initial approach.

Bus a_29 Power

The following figures show the power injection generated data and the actual data of bus a_29.

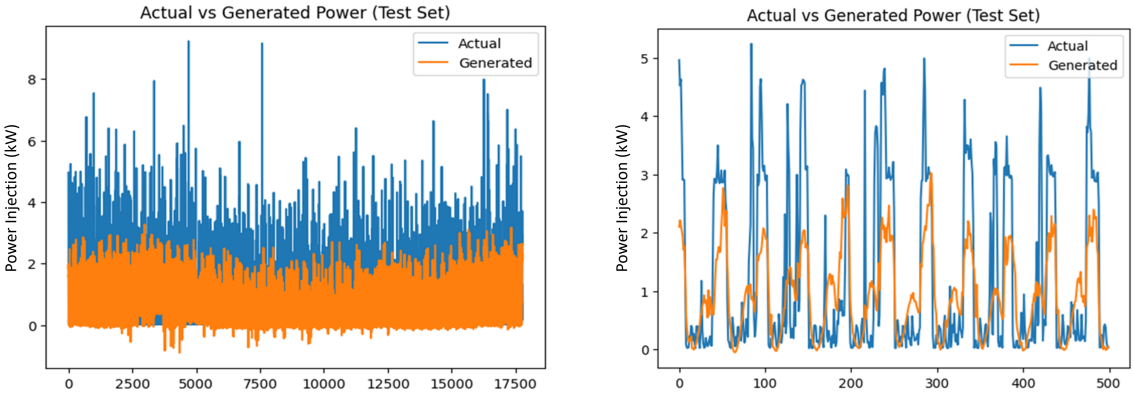


Figure 4.31: Power Injection Baseline Reconstruction Bus *a*_29

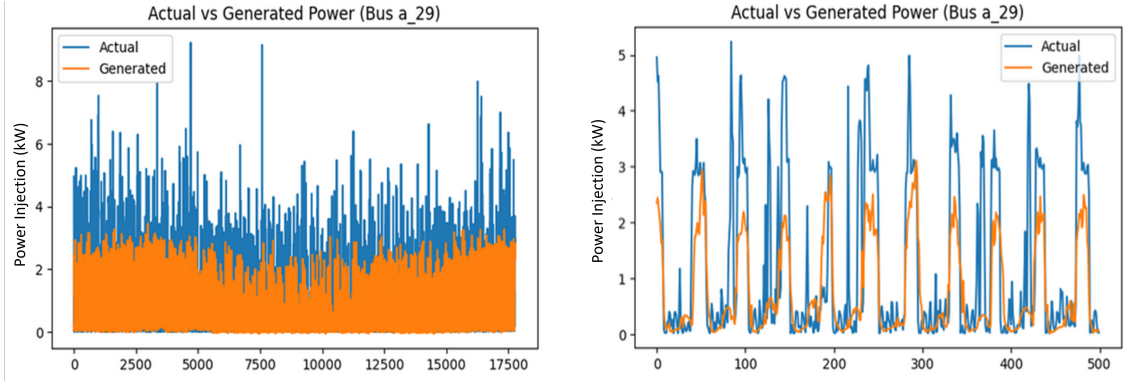


Figure 4.32: Power Injection SPPR Bus *a*_29

This bus has worse results than the one presented before; the numerical results and main conclusions are presented in subsection 4.6. Although, both reconstruction approaches exhibit very similar plots, as it is possible to see in Figures 4.31 and 4.32.

The first set of characteristics considered were the mean, standard deviation, skewness, kurtosis, and correlation.

Property	Real Values	Generated Values Baseline model	Generated Values SPPR model
Mean	0.745	0.759	0.609
Standard Deviation	1.075	0.629	0.764
Skewness	1.936	0.906	1.758
Kurtosis	3.657	0.589	2.010
Correlation	-	0.55	0.54

Table 4.29: Comparison of statistical properties between real and generated values from different models.

Several observations can be made based on comparing the mean, standard deviation, skewness, and kurtosis between the real and generated values.

The generated mean from the *Baseline* model (0.759208) is close to the real mean (0.745547), with a small difference between them. However, the mean from the *SPPR* model (0.60980) is notably lower than both the real mean and the *Baseline* model's generated mean.

The standard deviation of the generated values from the baseline model (0.629568) is noticeably lower than the real standard deviation (1.075725). The *SPPR* model's standard deviation (0.76470) is somewhat closer to the real standard deviation but still lower, indicating that the spread of values in both generated datasets is less than in the real data.

The skewness of the values generated by the baseline model (0.906411) is significantly lower than the real skewness (1.936946), suggesting a less skewed distribution. The *SPPR* model's skewness (1.75876) is closer to the real skewness but still lower, indicating that while the *SPPR* model improves on the *Baseline* model, it still does not fully capture the skewness of the real data.

The kurtosis of the baseline model's generated values (0.589403) is much lower than the real kurtosis (3.657676), showing a flatter distribution with thinner tails. The *SPPR* model's kurtosis (2.01044) is closer to the real kurtosis but still lower, reflecting a flatter distribution than the real data.

The correlation between the real and generated values from the *Baseline* model is 0.55, indicating a moderate positive linear relationship. The correlation for the *SPPR* model is slightly lower at 0.54, suggesting a similar linear trend but slightly decreased alignment.

Overall, while the generated data from both models approximates the real data reasonably well in terms of mean and correlation, there are notable differences in standard deviation, skewness, and kurtosis. The *SPPR* model offers improvements over the baseline model, particularly in skewness and kurtosis, but still falls short of accurately reproducing the full distribution characteristics of the real data. This indicates that further adjustments to the data generation process may be necessary to better capture these aspects of the real data.

A further method of elucidating the relationship between the actual and generated values is a detailed analysis of the relevant PDFs.

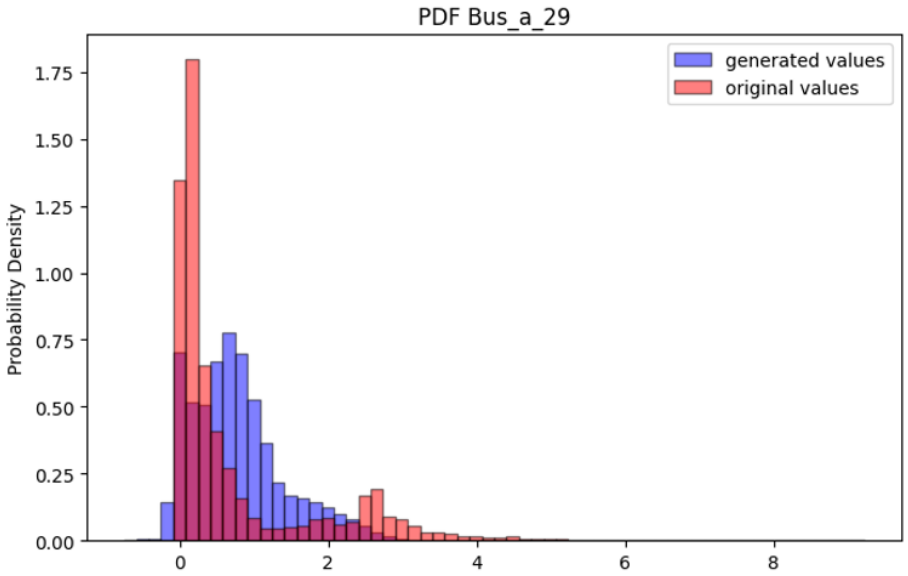


Figure 4.33: Baseline model Probability Density Function comparison Bus *a*_29

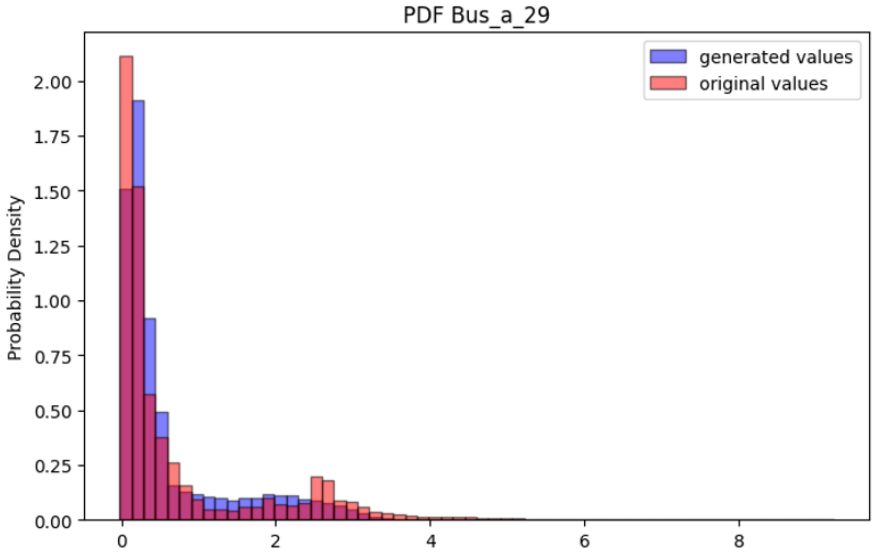


Figure 4.34: SPPR Probability Density Function comparison Bus *a*_29

The PDF comparison between the original and both generated series for Bus a_29 reveals key insights into the generation model's performance. The original values have a sharp peak near zero, indicating high concentration at lower values. In contrast, the generated values show a broader peak and greater variability on the baseline reconstruction case, while the SPPR case shows a closer behaviour with the original series with two peaks at zero and one, and it falls just like in the real values PDF.

The following analysis was conducted to analyze the autocorrelation function (ACF) and the partial autocorrelation function (PACF).

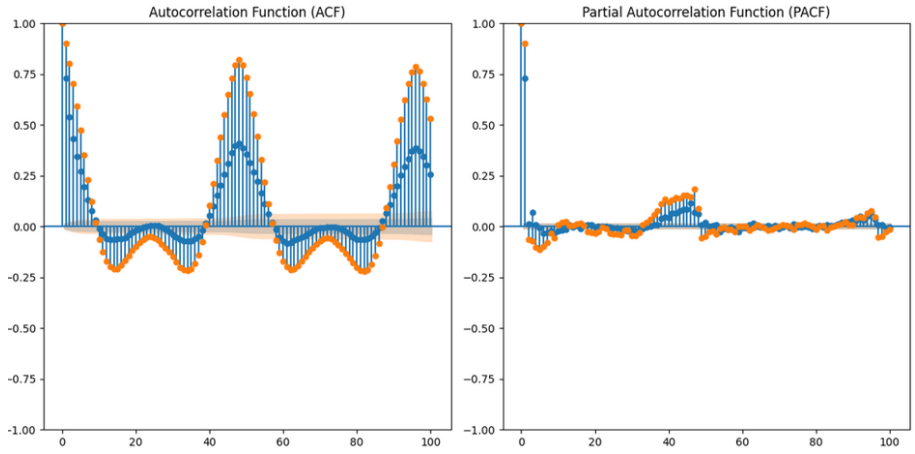


Figure 4.35: Baseline model ACF and PACF comparison Bus *a*_29

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots offer valuable insights into the properties of the original (blue) and generated (orange) time series. By examining these plots, we can compare how well the generated series replicates the characteristics of the original series. Starting with the ACF plots, the original and generated series exhibit significant peaks at regular intervals, approximately every 48 lags. This indicates that both series possess strong seasonality with a cycle length of about 48 observations. The original series (blue) displays pronounced peaks, suggesting strong and persistent autocorrelations at these intervals. The generated series (orange) mirrors this pattern, demonstrating that the seasonality has been successfully captured. Moving to the PACF plots, the original series shows a significant spike at lag 1, which is typical of an autoregressive process. This significant spike is also present in the generated series, indicating that the autoregressive behaviour has been effectively modelled. After the first lag, the PACF values for the original series hover around zero with some minor fluctuations, suggesting that the initial lags mainly capture the primary seasonality and autoregressive components. The generated series exhibits a similar pattern in the PACF plot, with values after the first lag also fluctuating around zero. While the overall pattern is comparable, the generated series displays minor variations in the magnitude of these fluctuations compared to the original series. The comparison of the ACF and PACF plots reveals that the generated series successfully replicates the key characteristics of the original series. The generated series captures the strong seasonality and the significant autoregressive component seen in the original series. However, the strength of the autocorrelations in the generated series is slightly weaker, as indicated by the lower amplitudes of the peaks in the ACF. Despite these minor differences, the overall patterns and behaviours are well preserved, demonstrating that the generated series approximates the original series well. This suggests that the methods used to generate the new series effectively model the underlying structure of the original time series.

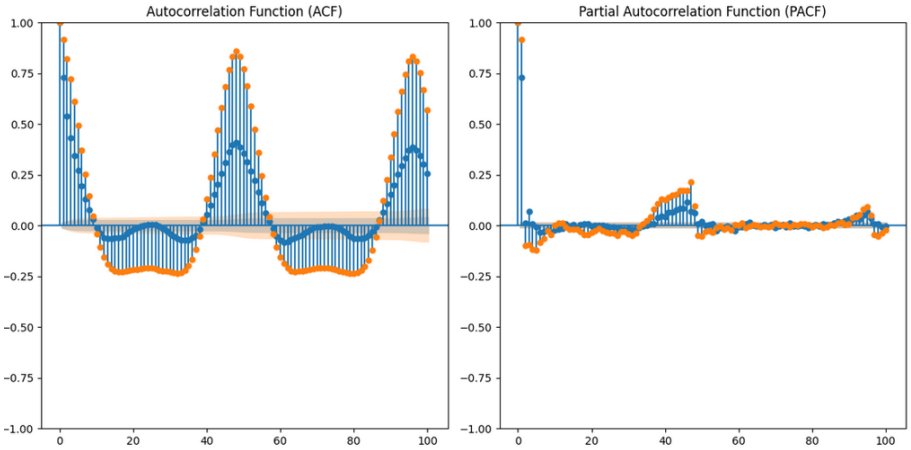


Figure 4.36: SPPR model ACF and PACF comparison Bus *a*_29

From the comparison of the autocorrelation functions (ACFs) obtained using the baseline model and the SPPR model, it can be concluded that the baseline model better replicates the negative peaks, indicating that the baseline model more effectively captures the relationships with previous lags compared to the SPPR model. However, it is important to note that the time series of this bus does not exhibit a strong relationship with previous lags, unlike the previous bus.

The next and final statistical analysis in this subsection analyzes the trend, seasonality, and residuals of the generated series compared to the real series.

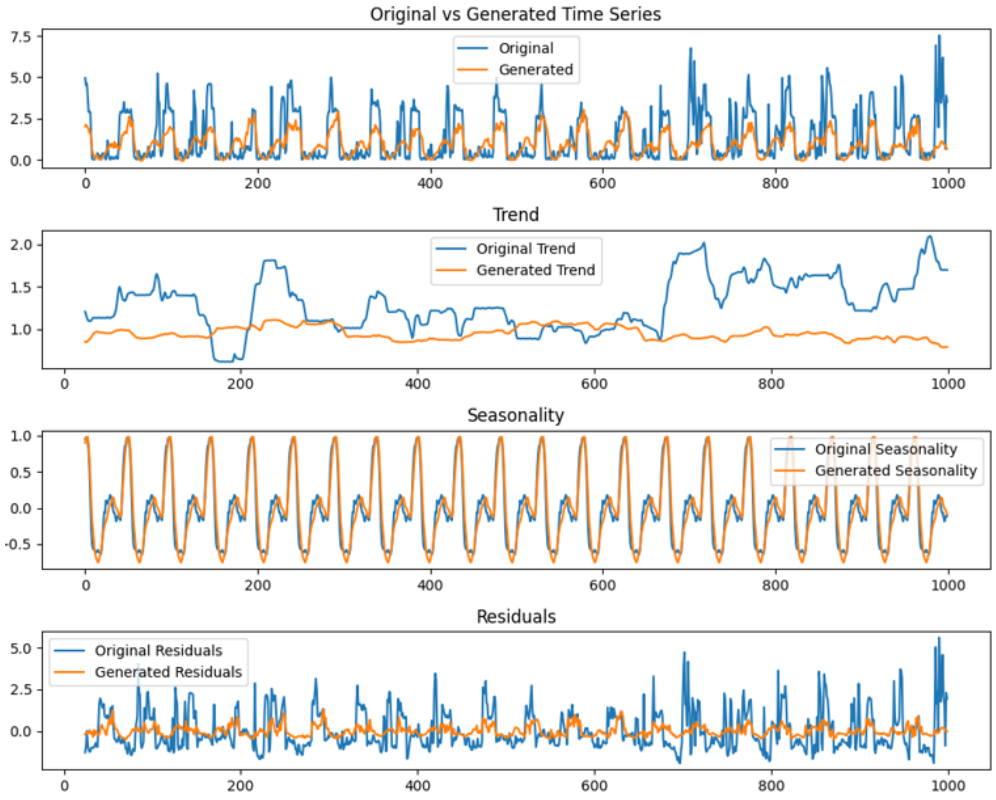


Figure 4.37: Baseline series decomposition Bus *a*_29

The first plot of Figure 4.37 compares the original and generated time series directly. The original series shows high variability with significant peaks and troughs, while the generated series appears to have similar fluctuations but with reduced amplitude. Although the generated series replicates the general pattern of the original, it does so with less intensity, suggesting a less volatile or more smoothed version of the original series.

The second plot of Figure 4.37 isolates the trend components of the original and generated series. The original trend displays considerable variation over time, with distinct upward and downward movements. In contrast, the generated trend is much flatter, showing minimal variation. This suggests that the model used to generate the new series may have difficulty capturing the long-term trends in the original data, leading to underestimating these components.

The third plot of Figure 4.37 focuses on the seasonal components of both series. The original and generated seasonality components display a clear, repeating pattern, indicating that the model has effectively captured the seasonality in the original data. The amplitude and frequency of the seasonal components in the generated series closely match those of the original, demonstrating that the seasonal characteristics are well-replicated.

After removing the trend and seasonality, the final plot of Figure 4.37 shows the residuals and the parts of the data remaining. The original residuals exhibit significant fluctuations, with occasional large spikes. While following a similar pattern, the generated residuals have lower amplitudes and are less volatile. This suggests that the generated series captures the general behaviour of the residuals but does not fully replicate the magnitude of the variations found in the original residuals.

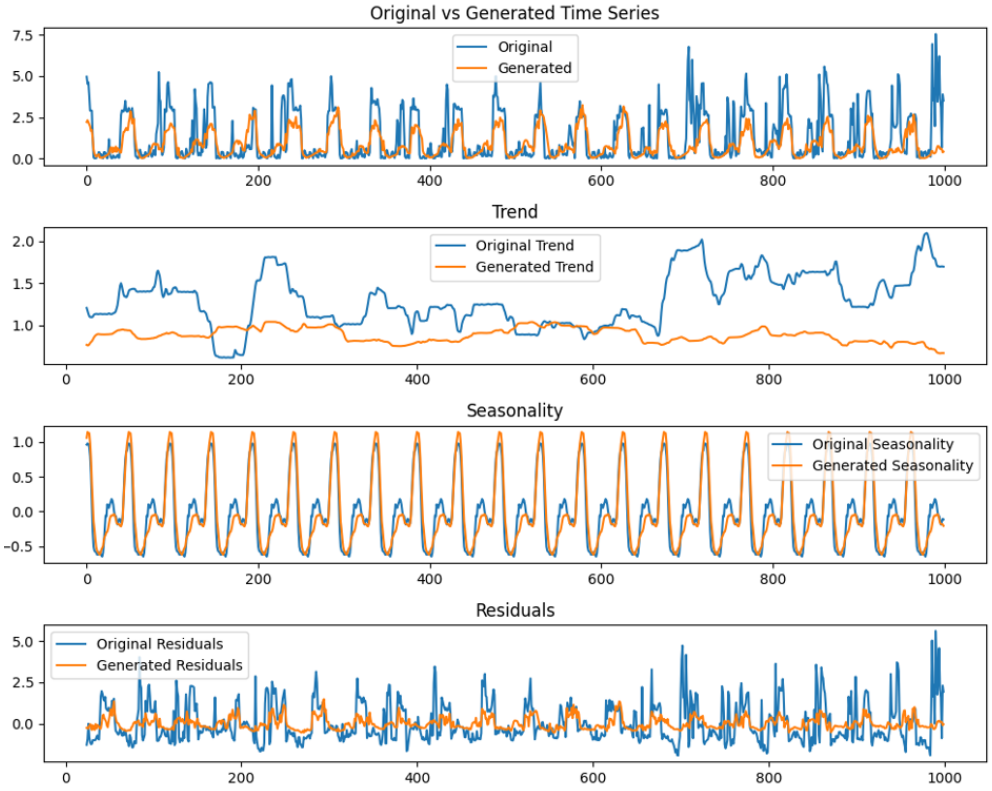


Figure 4.38: SPPR series decomposition Bus *a*_29

Both approaches effectively capture the general structure of the original time series, albeit with notable differences in their performance.

The baseline model demonstrates a generated time series that exhibits reduced amplitude and variability, which is particularly noticeable in the trend and residual components. This model follows the original series's overarching pattern but fails to replicate the dynamic fluctuations and peaks observed in the original data. The trend generated by the baseline model is relatively stable and lacks the intricate variability seen in the original trend. Similarly, the residuals produced by this model show some alignment with the original residuals but with significantly diminished variability and peak values.

In contrast, the *SPPR* model aligns more closely with the peaks of the generated time series and residuals than the baseline model. Although it does not fully capture the extent of variability in the original trend, the *SPPR* model demonstrates a slight improvement in reproducing the dynamic aspects of the time series. Both models capture the seasonality component well, aligning well with the original seasonal patterns.

Despite these advancements, neither the baseline nor the *SPPR* model achieves a perfect replication of the original time series. Each approach has its distinct strengths and weaknesses. The baseline model may be preferable in scenarios where a stable and less variable trend is sufficient. In contrast, the *SPPR* model could be more suitable when a closer alignment with peak values and dynamic fluctuations is required. The choice between these models ultimately hinges on the spe-

cific requirements of the task, such as the relative importance of accurately capturing seasonality versus the variability in the trend.

4.9 Physical properties Analysis

This section analyses the characteristics that relate smart meter data to the physical characteristics of the LV network under consideration. The main differences between the sensitivity coefficients calculated using real data and those calculated using generated data are shown. The first five buses of the network are presented to demonstrate the results.

	ΔP_{a_1}	ΔP_{a_2}	ΔP_{a_3}	ΔP_{a_4}	ΔP_{a_5}
ΔV_{a_1}	0	0	0	0	0
ΔV_{a_2}	0	-0.04683	0	-0.00021	0
ΔV_{a_3}	0	0	0	-0.00013	0
ΔV_{a_4}	0	0	0	-0.03211	0
ΔV_{a_5}	0	-0.04602	0	-0.00035	0

Table 4.30: Sensitivity Coefficients for Real data time series (First Five buses)

	ΔP_{a_1}	ΔP_{a_2}	ΔP_{a_3}	ΔP_{a_4}	ΔP_{a_5}
ΔV_{a_1}	0	0	0	0	0
ΔV_{a_2}	0	-0.03668	0	0	0
ΔV_{a_3}	0	0	0	0	0
ΔV_{a_4}	0	0	0	-0.02950	0
ΔV_{a_5}	0	-0.03039	0	0	0

Table 4.31: Sensitivity Coefficients for Generated data time series (First Five buses)

The analysis reveals that most sensitivity coefficients between changes in active power (ΔP) and voltage (ΔV) across buses are zero for both real and generated data because these buses have no injected power. This suggests that in this LV network, alterations in power consumption at most buses do not significantly impact voltage levels at other buses. Such a pattern indicates a stable network where local power changes do not propagate widely through voltage fluctuations.

However, where non-zero sensitivity coefficients occur, there are slight differences in magnitude between the real data and the generated data. For instance, coefficients like $\Delta V_{a_2} \rightarrow \Delta P_{a_2}$ and $\Delta V_{a_5} \rightarrow \Delta P_{a_2}$, and $\Delta V_{a_4} \rightarrow \Delta P_{a_4}$, exhibit small variations in their values between the two datasets. This suggests that while the generated data broadly mirrors the real network behaviour, nuances in sensitivity may arise from discrepancies in the data generation process or underlying assumptions.

Specific observations highlight that certain buses, such as bus 2 and bus 4, consistently show non-zero sensitivities, indicating potential areas of interest for network analysis. For instance, ΔP_{a_2} affects ΔV_{a_2} in both datasets, albeit with slight differences in impact strength. Meanwhile,

ΔP_{a_4} influences ΔV_{a_2}, ΔV_{a_3}, ΔV_{a_4} and ΔV_{a_5} in the real data, but only ΔV_{a_4} in the generated data, indicating a nuanced response that may require further investigation into network topology and load distribution.

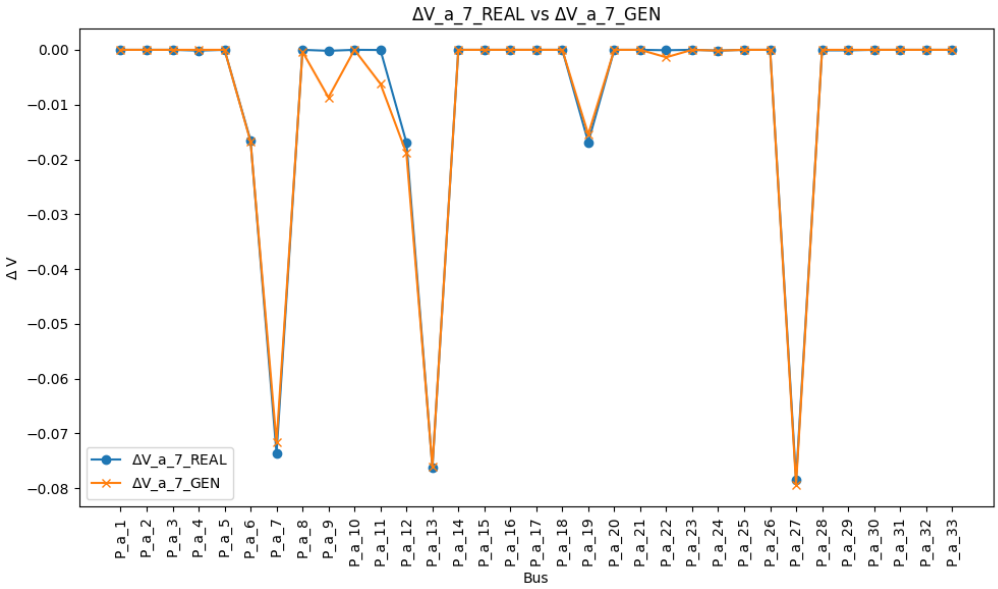


Figure 4.39: Sensitivity coefficients analysis for bus a_7

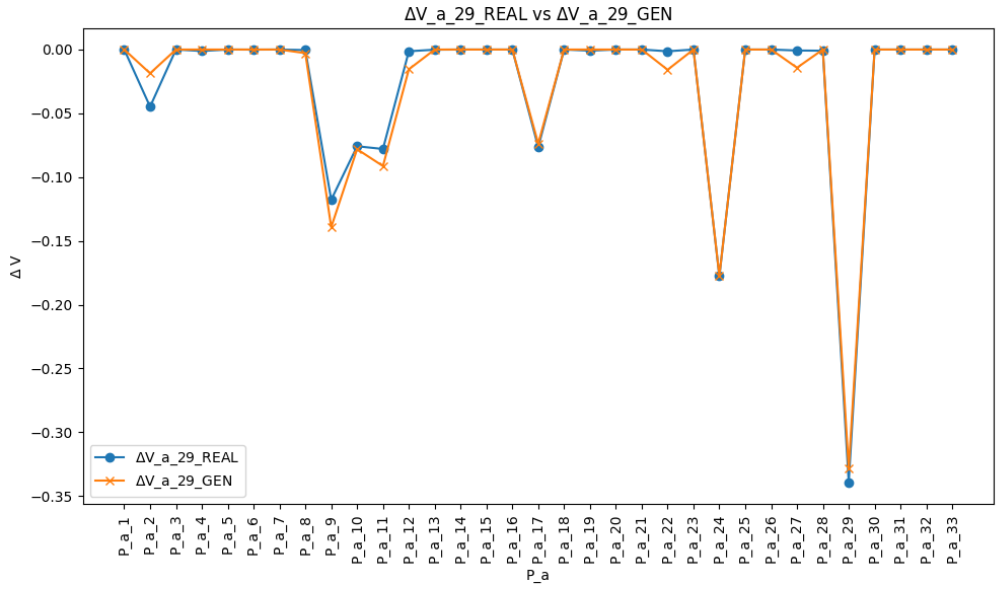


Figure 4.40: Sensitivity coefficients analysis for bus a_29

Looking at the results of Figure 4.39 and Figure 4.40 , which correspond to the sensitivity coefficients of buses 7 and 29, it can be seen that, despite the small variations in the magnitude of some of the powers, their behaviour is very close to the real data for the data generated, which allows us to conclude that the physical properties of the network are generally maintained.

The consistency between real and generated data overall supports the utility of generated data for approximating network behaviour. However, the observed discrepancies underscore the importance of validating generated data against real-world measurements to ensure accuracy in network analyses and decision-making processes. These findings suggest that while generated data can provide a general understanding of network dynamics, caution should be exercised in relying solely on generated data for critical assessments and planning.

In conclusion, while generated data serves as a valuable tool for approximating real network behaviour, the study emphasizes the necessity of combining it with real-world data to achieve comprehensive insights into network performance and resilience. This integrated approach ensures robust network planning and management strategies in practical applications.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The research has demonstrated significant advancements in reconstructing power injection and voltage time series by integrating various data sources and applying different modelling techniques.

In terms of voltage time series reconstruction, the results of the study indicate that Linear Regression (LR) models, particularly those incorporating real-time substation voltage data, achieved superior performance in terms of both accuracy and computational efficiency, with the LR_2 and LR_3 models showing the smaller RMSE and MAE, these models also maintained favourable statistical properties, closely matching the actual stress data's mean, standard deviation, skewness and kurtosis. LSTM models proved advantageous when continuous real-time data was unavailable, as they effectively retained information over extended periods and exhibited realistic seasonality and trend components. In addition, including substation voltage data in the single-phase voltage reconstruction significantly improved the prediction accuracy. The statistical properties, such as correlation and autocorrelation, showed that these models could effectively capture temporal dependencies and relationships within the data, making them robust tools for voltage time series reconstruction in power grids.

Integrating PV and EV data for power injection time series reconstruction significantly improved model accuracy and robustness. LR and LSTM models were validated as practical methods, with LSTMs excelling in handling data complexity and non-linearities. The reconstructed power injection time series retained accurate statistical properties, including mean, standard deviation, skewness, kurtosis and autocorrelation, ensuring a close resemblance to real-world data. This comprehensive approach to incorporating diverse data sources resulted in models that not only met performance metrics but also preserved the statistical integrity of the power injection data, highlighting their practical applicability for improving grid reliability and management.

5.2 Future Work

To build on these findings, several future research directions and practical implementations are recommended:

- **Incorporation of additional data sources**: Future studies could explore integrating additional data sources such as weather data, consumer behaviour analytics, and other environmental factors further to enhance model accuracy, robustness, and statistical fidelity.
- **Advanced machine learning techniques**: Exploring more advanced machine learning techniques, including deep learning, hybrid models, and ensemble learning, could improve accuracy and robustness. Techniques like transfer learning might also be beneficial for adapting models to new datasets without extensive retraining, ensuring statistical properties are preserved.
- **User behaviour analysis**: Incorporating user behaviour analysis could help understand how consumer actions impact energy consumption and production patterns. This could lead to more personalised and efficient energy management solutions, enhancing grid reliability and consumer satisfaction while preserving the statistical properties of the data.
- **Rare Events Modeling**: Another future critical study is to represent rare events within the reconstructed historical dataset accurately. Due to their infrequency, rare events challenge conventional data reconstruction.

Bibliography

[1] I. B. F. Marques, "Impact of smart meter data availability in data-driven low voltage management," Master's Thesis, FEUP, UP, 2022. [Online]. Available: https://repositorio-aberto.up.pt/bitstream/10216/144542/2/587606.pdf.

[2] C. Chalmers, W. Hurst, M. Mackay, and P. Fergus, "Smart health monitoring using the advance metering infrastructure," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015, pp. 2297–2302. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.339.

[3] B. Völker, A. Reinhardt, A. Faustine, and L. Pereira, "Watt's up at home? smart meter data analytics from a consumer-centric perspective," *Energies*, vol. 14, no. 3, 2021, ISSN: 1996-1073. DOI: 10.3390/en14030719. [Online]. Available: https://www.mdpi.com/1996-1073/14/3/719.

[4] I. f. E. Joint Research Centre, Transport, P. Zangheri, P. Bertoldi, and T. Serrenho, "Energy feedback systems – evaluation of meta-studies on energy savings through feedback – energy efficiency directive articles 9-11 on feedback, billing and consumer information," Publications Office, Tech. Rep., 2016, https://data.europa.eu/doi/10.2790/565532.

[5] Zangheri, Serrenho, and Bertoldi, "Energy savings from feedback systems: A meta-studies' review," *Energies*, vol. 12, no. 19, p. 3788, Oct. 2019. DOI: 10.3390/en12193788.

[6] W. Dai, X. Liu, A. Heller, *et al.*, *Smart Meter Data Anomaly Detection Using Variational Recurrent Autoencoders with Attention*. Springer Nature, Jan. 2022.

[7] A. Santolamazza, V. Cesarotti, and V. Introna, "Anomaly detection in energy consumption for condition-based maintenance of compressed air generation systems: An approach based on artificial neural networks," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1131–1136, 2018. DOI: 10.1016/j.ifacol.2018.08.439.

[8] M. Afzalan and F. Jazizadeh, "Residential loads flexibility potential for demand response using energy consumption patterns and user segments," *Applied Energy*, vol. 254, Art. 113693, 2019, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.113693.

[9] H. Çimen, N. Çetinkaya, J. C. Vasquez, and J. M. Guerrero, "A microgrid energy management system based on non-intrusive load monitoring via multitask learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 977–987, Mar. 2021. DOI: 10.1109/TSG.2020.3027491.

[10] M. Pirbazari, A. Mehdipour, M. Farmanbar, A. Chakravorty, and C. Rong, "Short-term load forecasting using smart meter data: A generalization analysis," *Processes*, vol. 8, no. 4, p. 484, 2020. DOI: 10.3390/pr8040484.

[11] J. T. Lee, J. Freitas, I. L. Ferrall, D. M. Kammen, E. Brewer, and D. S. Callaway, "Review and perspectives on data sharing and privacy in expanding electricity access," *Proceedings of the IEEE*, vol. 107, no. 9, pp. 1803–1819, 2019. DOI: 10.1109/JPROC.2019.2919306.

[12] S. Pfenninger, J. DeCarolis, L. Hirth, S. Quoilin, and I. Staffell, "The importance of open data and software: Is energy research lagging behind?" *Energy Policy*, vol. 101, pp. 211–215, 2017, ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2016.11.046. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421516306516.

[13] W. Medjroubi, U. P. Müller, M. Scharf, C. Matke, and D. Kleinhans, "Open data in power grid modelling: New approaches towards transparent grid models," *Energy Reports*, vol. 3, pp. 14–21, 2017, ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2016.12.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484716300877.

[14] A. Pinceti, L. Sankar, and O. Kosut, "Generation of synthetic multi-resolution time series load data," *IET Smart Grid*, vol. 6, no. 5, pp. 492–502, 2023. DOI: https://doi.org/10.1049/stg2.12116. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/stg2.12116. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/stg2.12116.

[15] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3258–3265, 2017. DOI: 10.1109/TPWRS.2016.2616385.

[16] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1502–1510, 2017. DOI: 10.1109/TPWRS.2016.2586460.

[17] E. Schweitzer and A. Scaglione, "A mathematical programming solution for automatic generation of synthetic power flow cases," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 729–741, 2019. DOI: 10.1109/TPWRS.2018.2863266.

[18] H. Nfaoui, H. Essiarab, and A. Sayigh, "A stochastic markov chain model for simulating wind speed time series at tangiers, morocco," *Renewable Energy*, vol. 29, no. 8, pp. 1407–1418, 2004, ISSN: 0960-1481. DOI: https://doi.org/10.1016/S0960-1481(03)00143-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148103001435.

[19] G. Papaefthymiou and B. Klockl, "Mcmc for wind power simulation," *Energy Conversion, IEEE Transactions on*, vol. 23, pp. 234–240, Apr. 2008. DOI: 10.1109/TEC.2007.914174.

[20] R. Palma-Behnke, F. Valencia, J. Vega-Herrera, and O. Núñez-Mata, "Synthetic time series generation model for analysis of power system operation and expansion with high renewable energy penetration," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 849–858, 2021. DOI: 10.35833/MPCE.2020.000747.

[21] R. Aguiar and M. Collares-Pereira, "Tag: A time-dependent, autoregressive, gaussian model for generating synthetic hourly radiation," *Solar Energy*, vol. 49, no. 3, pp. 167–174, 1992, ISSN: 0038-092X. DOI: https://doi.org/10.1016/0038-092X(92)90068-L. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0038092X9290068L.

[22] X. Zhou, C. Liu, Y. Luo, *et al.*, "Wind power forecast based on variational mode decomposition and long short term memory attention network," *Energy Reports*, vol. 8, pp. 922–931, 2022, 2022 The 5th International Conference on Electrical Engineering and Green Energy, ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2022.08.159. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235248472201602X.

[23] A. Lojowska, D. Kurowicka, G. Papaefthymiou, and L. Sluis, "Advantages of arma-garch wind speed time series modeling," Jul. 2010, pp. 83–88. DOI: 10.1109/PMAPS.2010.5528979.

[24] Z. De Greve, C. Daniels, L. Exizidis, F. Vallée, and J. Lobry, "A review of time series models for the long term modeling of wind speed in distribution network planning," in *CIRED Workshop*, Paper No. 0363, Rome, Jun. 2014, pp. 1–5.

[25] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi, "A bottom-up approach to residential load modeling," *IEEE Transactions on Power Systems*, vol. 9, no. 2, pp. 957–964, 1994. DOI: 10.1109/59.317650.

[26] N. Hodzic, T. Konjic, and V. Miranda, "Artificial neural networks applied to short term load diagram prediction," in *2006 8th Seminar on Neural Network Applications in Electrical Engineering*, 2006, pp. 219–223. DOI: 10.1109/NEUREL.2006.341217.

[27] T. A. C. Guimarães, "Load forecast on a micro grid level through machine learning algorithms," https://hdl.handle.net/10216/127087, M.S. thesis, Faculdade de Engenharia, Universidade do Porto, Apr. 2020.

[28] N. Iftikhar, X. Liu, S. Danalachi, F. E. Nordbjerg, and J. H. Vollesen, "A scalable smart meter data generator using spark," in *On the Move to Meaningful Internet Systems. OTM 2017 Conferences - Confederated International Conferences: CoopIS, C and TC, and ODBASE 2017, Proceedings*, H. Panetto, A. Paschke, R. Meersman, *et al.*, Eds., Springer, 2017, pp. 21–36. DOI: 10.1007/978-3-319-69462-7_2.

[29] S. Park, H. Kim, H. Moon, J. Heo, and S. Yoon, "Concurrent simulation platform for energy-aware smart metering systems," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1918–1926, 2010. DOI: 10.1109/TCE.2010.5606347.

[30] E. P., R. L., X. Liu, and J. B. Holm-Nielsen, "Smart meter synthetic data generator development in python using fbprophet," *Software Impacts*, vol. 15, p. 100 468, 2023, ISSN: 2665-9638. DOI: https://doi.org/10.1016/j.simpa.2023.100468. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963823000052.

[31] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Generative adversarial network for synthetic time series data generation in smart grids," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2018, pp. 1–6. DOI: 10.1109/SmartGridComm.2018.8587464.

[32] S. Asre and A. Anwar, "Synthetic energy data generation using time variant generative adversarial network," *Electronics*, vol. 11, no. 3, 2022, ISSN: 2079-9292. DOI: 10.3390/electronics11030355. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/355.

[33] T. Hong, D. Macumber, H. Li, and et al., "Generation and representation of synthetic smart meter data," *Building Simulation*, vol. 13, no. 6, pp. 1205–1220, 2020. DOI: 10.1007/s12273-020-0661-y.

[34] G. D. Grosso, G. Pichler, and P. Piantanida, *Privacy-preserving synthetic smart meters data*, 2021. arXiv: 2012.04475 [eess.SP].

[35] A. Biswal, *Recurrent neural network (rnn) tutorial: Types and examples [updated]*, https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn, Nov. 2022.

[36] J. Brownlee, *A gentle introduction to long short-term memory networks*, https://machinelearningma com/gentle-introduction-long-short-term-memory-networks-experts/, Accessed on: June 3, 2024.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[38] G. Sampaio, R. J. Bessa, C. Goncalves, and C. Gouveia, "Conditional parametric model for sensitivity factors in lv grids: A privacy-preserving approach," *Electric Power Systems Research*, vol. 211, p. 108 316, 2022.

[39] R. J. Bessa, A. Trindade, and V. Miranda, "Spatial-temporal solar power forecasting for smart grids," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 232–241, 2015. DOI: 10.1109/TII.2014.2365703.

[40] Commission for Energy Regulation (CER), "Electricity smart metering customer behaviour trials findings report," Commission for Energy Regulation, Dublin, Ireland, Technical Report CER/11/080a, 2011.

[41] J. D. Cross and R. Hartshorn, "My electric avenue: Integrating electric vehicles into the electrical networks," in *6th Hybrid and Electric Vehicles Conference (HEVC 2016)*, 2016, pp. 1–6. DOI: 10.1049/cp.2016.0972.