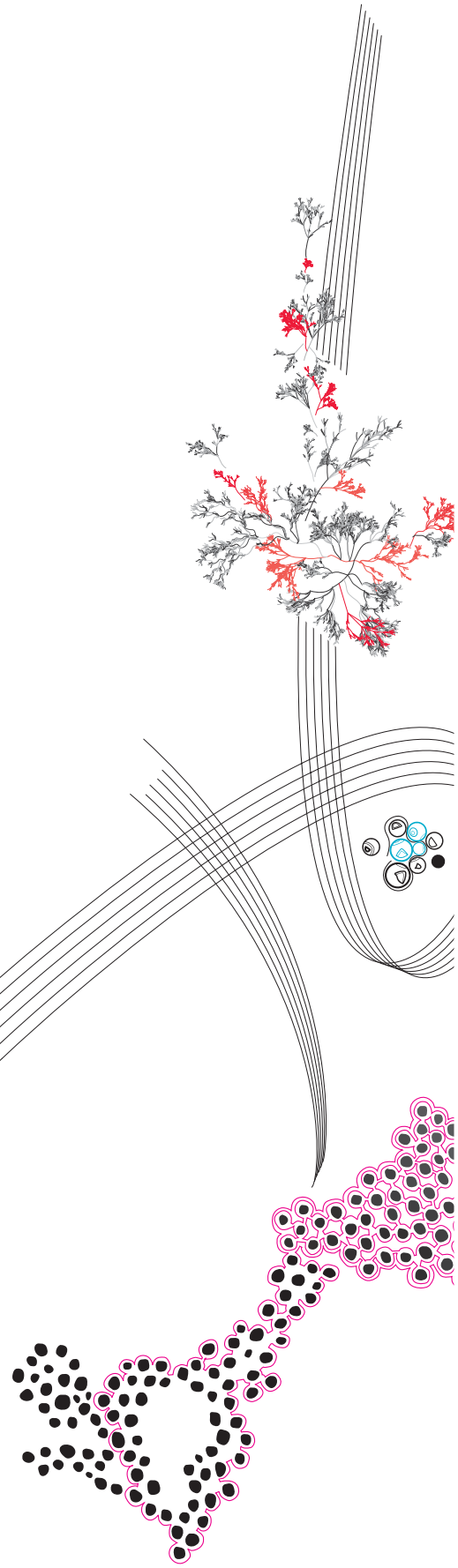

BSc Creative Technology

Graduation Project

A Comparative Study of Explainable AI Techniques for Short-Term Energy Forecasting: Evaluating WindowSHAP and TimeSHAP across Time-series Prediction Models

Koki Omura

July, 2025

Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

Supervisor: Dr. Faizan Ahmed, Dr. Deepak Tunuguntla
Critical observer: MSc. Annemarie Jutte

UNIVERSITY OF TWENTE.

Abstract

This paper analyses the compatibility of the different explanation models for time-series load forecasting by evaluating them based on the quantitative method. The aim of the analysis is inspired by the digital twin development project for efficient energy distribution of Ecofactorij, in which the energy load is expected to be accurately forecasted using AI tools. However, the current prediction models are often described as black-box, which means the model cannot project insights into its own decision-making process for the outputs. In order to solve this, several explanation methods have been developed and deployed to address this lack of transparency. Therefore, this analysis aims to reveal the advantages, tendencies, and limitations of each combination of an explanation model and a prediction model for time-series forecasting. In particular, this paper argues for WindowSHAP and TimeSHAP, the two different variants of Shapley Additive Explanation(SHAP) by quantitatively scoring them based on their fidelity, stability, and sparsity to examine the characteristics of each coupling in combination with a time-series prediction model with a black-box problem, namely Recurrent Neural Network(RNN), Long Short-term Memory(LSTM), and Gated Recurrent Unit (GRU). A human-subject Likert scale survey was used to qualitatively evaluate each explanation method's performance, which was also evaluated using fidelity, sparsity, and stability metrics. While both approaches show limitations in stability and user interpretability, the results show that TimeSHAP performs slightly better than WindowSHAP across all metrics, especially fidelity. The results support the continuous efforts towards transparent AI systems for sustainable energy management and provide useful insights into the application of XAI in time-series forecasting.

Keywords: Explainable AI, WindowSHAP, TimeSHAP, Time-Series Forecasting, Digital Twin, Energy Load Forecasting

Acknowledgement

I would like to express my deepest and heartfelt appreciation for Dr. Ahmed Faizan, Dr. Deepak Tunuguntla, and MSc. Annemarie Jutte for their helpful and valuable guidance throughout this project with their patience, precious time and effort. The resources and educational insights received from them have been significantly instrumental in the completion of the project. Furthermore, I wish to extend my sincere gratitude to any other stakeholders who contributed to the conduct of this project with unwavering respect and kindness.

Contents

1 Introduction

This project focuses on achieving an effective utilisation of Artificial Intelligence (AI)/machine learning(ML) models in time-series forecasting. The project is inspired by Ecofactorij [1], a sustainable business park located in Apeldoorn, the Netherlands. Ecofactorij envisions the utilisation of AI/ML technologies to enable efficient and effective energy provision to each facility. In this context, a technique called time-series forecasting is a relevant topic. Time-series forecasting is a data-driven statistical approach that makes predictions and informs strategies in the decision-making process [2]. There are several models to achieve this, namely, Long Short-Term Memory (LSTM), XGBoost, Facebook Prophet and so forth. Speaking of current challenges in time-series forecasting, there are examples of generalizability and accuracy. As their primary role suggests, a certain number of research studies regarding the enhancement of model accuracy and generalisability have been conducted. However, in the interest of achieving effective forecasting, it is undeniable that some of the current prediction models are unable to provide reasoning behind their decisions. Brożek et al. [3] describe this opacity issue by the name of the "black box" problem. The black box problem refers to the model's limited ability to explain the patterns and the algorithms upon which the output relies [3]. The black box problem involves inherently multilateral problems. First of all, with a lack of a model's explainability, the model is unable to provide the stakeholders with proper information about the justified reasoning for the leading conclusion. This is critical because without the information, the model cannot be properly assessed, dismissing potential bias and failing to provide the developers access to a transparent model tuning in the process of debugging or enhancement. Within this context, the concept of Explainable AI (XAI) has arisen to address the problem. XAI methods were developed to make it possible for us to understand the complexity behind model decisions by providing the path on which the model came to a conclusion [4]. Although a certain number of studies have been conducted to tackle this issue by implementing explanation tools such as Shapley additive explanations(SHAP), it appears to be yet underexplored in terms of studies concerning the applications of these tools to time-series forecasting models. Therefore, this study will introduce the two prominent sequential specific model variants of SHAP, TimeSHAP and WindowSHAP, assessing their effectiveness in generating short-term interpretable explanations using the real-world dataset collected from Ecofactorij in 2024.

2 Literature Review

This chapter reviews the existing research and studies relevant to the application of SHAP methods. The overarching aim of this chapter is to frame the research questions introduced in the previous chapter into a deeper theoretical and methodological context, thereby establishing a foundation for the subsequent experimental analysis.

2.1 Relevant Research on Explanation Methods

2.1.1 Explanation methods for time-series models

In 2025, A.M. Salih et al. [5] have listed some of the explanation tools that have been popular among the Github community and sorted them into a visualised graph based on the Github review stars, which implies that SHAP and LIME are the current major explanation tools that have seen increasing adoption. Nevertheless, both SHAP and LIME have been largely employed in miscellaneous utilisation contexts across fields. In 2022, T. B. Çelik et al. proposed integrating LIME into financial time series prediction, in which they claimed that the use of LIME was effective, increasing its prediction accuracy by proposing their original integration framework. They explored the whole through methods to improve its accuracy by decomposing the process into multiple phases, starting from preprocessing the time-series dataset using the technique called Empirical Mode Decomposition (EMD) and predicting them on an Artificial Neural Network(ANN) and then Random Forest(RF) models and finally applying LIME to ensure its coherency of the expected output. This study suggests a strong compatibility with the time-series prediction of LIME. Another research conducted by Zhang et al. [6], which is highly relevant to this project based on its dataset comparing solar power energy generation and its consumption amount, analysed five different prediction models consisting of M5 Gradient Boosting Machine (LightGBM), RF, Bidirectional Recurrent Neural Network (Bi-RNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Gated Recurrent Unit (Bi-GRU) models using the explanation method SHAP for feature contribution. They conclude that M5 LightGBM outperformed the other prediction models in its robustness and efficiency, and SHAP enhanced its transparency. Aside from methods like SHAP and LIME, research such as that from Xinli Yu et al. [7] explores an explainable approach using Large language models(LLM), like GPT-4 and Open LLaMA, confirming that it is effective when manipulating heterogeneous datasets that require an understanding of both numerical and textual information.

2.1.2 Comparative Studies on SHAP

One of the latest studies by Ahmed et al. [8] conducted a comparative analysis on SHAP and LIME in the detection of diabetes based on a survey using ML models like Logistic regression(LR) and RF. The papers figured out that both SHAP and LIME showed a signal of model-dependent tendency. The interesting thing here is the fact that they found such characteristics in LIME, although LIME is commonly known as model-independent or model-agnostic, meaning that LIME can be applied to any black-box classifier or regressor without the need to access the internal structure of the prediction models [9]. Another study by Raufi et al. [10] suggests that SHAP showed superior performance in combination with an ANN model in detecting credit card fraud. Furthermore, Mane et al. [4] comprehensively compared SHAP and LIME using simple machine learning models like SVM, RF and decision tree models. Their study claims that each of the tools differs in its strengths and weaknesses, providing key insights on future implementation that the combination of these two tools would strengthen the outcomes and improve the overall model transparency.

Likewise, Mitra and Gilpin [11] claim that the functionality of these explanation methods is affected by the dataset and the prediction model types. Finally, research on this comparison in time-series forecasting was conducted by [12], which used these tools for forecasting sales deals for a consultancy company and evaluated their explainability based on a human survey, resulting in LIME providing better performance with a slightly higher score than SHAP, if not a landslide. They conclude that LIME is capable of providing meaningful insights into the decision-making process and has accumulated more trust from users. While both SHAP and LIME are extensively functional in a wide spectrum of interpretability scenarios, comparing these two methods requires a substantial amount of background knowledge and techniques of feature engineering, statistics, and ML engineering.

2.2 TimeSHAP and WindowSHAP

While SHAP alone is capable of providing how much SHAP values are distributed to each timepoint in a sequence, conventional SHAP requires a substantial calculation cost and computing performance. Given such context, TimeSHAP and WindowSHAP emerged to mitigate these constraints. Although both TimeSHAP and WindoSHAP share some similarities in their challenge to be addressed, their approaches diverge. TimeSHAP was primarily developed as an extension of KarnelSHAP, appending a pruning function that retains only a significant part of the input sequence [13]. TimeSHAP also makes it accessible for the stakeholders to perceive the SHAP value distribution at the timestep-level and variable-level. By the nature of pruning techniques, it became more intuitive to discern the exact causality between the prediction output and features. On the other hand, WindowSHAP accelerates its SHAP calculation by grouping timepoints as a window partition while KernelSHAP treats each timepoint individually [14]. It led to a greater plain interpretation by capturing the timesteps as a rather long-term trend in the sequence, which becomes evident when the distributed SHAP values exhibit consecutive fluctuations.

2.3 Preliminary Findings

With an emphasis on SHAP and LIME—two popular and flexible XAI approaches—this chapter examined a broad range of explanation tools pertinent to time-series forecasting. Although both approaches have advantages, comparative studies reveal that the prediction model and dataset properties frequently determine how effective a method is. Despite being referred to as model-agnostic, LIME can still exhibit model-dependent behaviour, according to recent research.

There are still few systematic comparisons of SHAP variants, especially TimeSHAP and WindowSHAP, despite the growing significance of explainability in time-series forecasting. Although these modifications were intended to solve computational issues and improve the interpretability of sequential data, their applicability and practical performance—particularly in energy forecasting—have not been fully investigated.

The need for both technical and human-centered evaluations is further highlighted by the fact that many current studies ignore how human users perceive and interpret explanations. In order to bridge that gap and promote more interpretable AI in time-series applications, this study compares

TimeSHAP and WindowSHAP using quantitative metrics.

2.4 Research Questions

The comprehensive aim of this study is to explore how WindowSHAP and TimeSHAP differ in their value distribution nature and trend in a time sequence when applied to the same prediction target point. To structure this analysis, the deployment of a synced plot visualisation is necessary for a fair and clear evaluation.

2.4.1 RQ1: How do explanation techniques (WindowSHAP vs. TimeSHAP) differ in performance when applied to time-series energy demand Predictions?

This question aims to support the idea of the project delving into discovering how and whether WindowSHAP and TimeSHAP outperform one another in the nature of SHAP calculation of each time point in the sequence. The answer to this question would reveal their effectiveness in each specific use case and data background context in further implementation.

2.4.2 RQ2: How can explanation methods be improved to enhance stakeholder understanding of how much each timestep contributed to the prediction output?

WindowSHAP and TimeSHAP are customised to visualise the SHAP calculation results differently from one another in their plot methods. While WindowSHAP shows how much each timepoint in a sequence contributed to calculating the prediction result, TimeSHAP aims to emphasise how much each of the independent variables affected the prediction result. Therefore, it is not feasible to compare these two tools directly, but rather requires synchronisation in their visualisation style. Consequently, this modification would result in achieving simpler and clearer visualisation for stakeholders.

3 Methodology

3.1 CRISP-DM

Throughout the study, this project follows a structured framework of conduct. Nonetheless, the paper decided to apply the Supervisor-recommended methodology called the Cross-Industry Standard Process for Data Mining (CRISP-DM). This methodology is meant to be applied as a form of guideline that enables the greater efficiency and effectiveness of the conduct. For more concrete details, the manual provided by Data Science PM [15] will be referred to. CRISP-DM consists of six iterative stages: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*. These stages guide the overall structure and workflow of the research process.

This paper applies CRISP-DM as follows:

- **Business Understanding**: Enabling effective and efficient energy forecasting at Ecofactorij, by providing the ML model that is suitable for the achievement of the digital twin of the energy grids at the business park.
- **Data Understanding**: A Dataset of the historical energy consumption will be provided. The data includes the consumption amount of renewable energy.
- **Data Preparation**: This phase includes preprocessing of the dataset for application onto the prediction models and WindowSHAP and TimeSHAP.
- **Modelling**: Running the dataset through the selected prediction models and the explanation models. This phase may include the adjustment of the models to explicitly visualise the explanation outputs.
- **Evaluation**: Assessing the model compatibility of each coupling, revealing its strengths and weaknesses based on quantitative evaluation based on the metrics.
- **Deployment (none)**: The practical deployment of any methods will not be involved within the scope of this project; instead, potential future implementation will be discussed from the results.

3.2 Business Understanding

3.2.1 Objectives

The objective of this project is to compare the two different explanation methods, WindowSHAP and TimeSHAP, against time-series forecasting models. However, in order to conduct the study, the project requires some form of objective indicator for the assessment of these tools. Therefore, the project will employ quantitative evaluation criteria based on the quality of the method's functionality. These can be done through a comprehensive analysis of each coupling of the prediction model and the explanation methods. Consequently, the project would perform a human survey to assess whether the generated explanatory outputs are intuitively interpretable to human users. One of the stakeholders in this project is Ecofactorij, which intends to construct a smart grid system. However, at the current stage of development, they are planning to experiment with the energy grids in the digital twin(DT) that emulates the energy distribution of the facilities in high fidelity, enabling precise simulation and optimisation. For the DT, an AI/ML model that can process the historical data of the energy demand is required. The ultimate goal of Ecofactorij's project is to enable the assessment of whether the factory is making the best decision according to the criteria that consist of (i) earnings on the various energy markets; (ii) maximum use of generated sustainable energy (solar power); (iii) robustness of the grid; and (iv) utilisation rate of batteries and the network. In order to visualise this information to the stakeholders, the fixation of the interpretability issue seems essential. The author of this paper is from the University of Twente and is conducting a technical comparison of the explanation tools. The eventual goal of this project is an investigation of WindowSHAP and TimeSHAP, the two optimal explanation tools, on their

effectiveness in practical usage in combination with time-series prediction models. Eventually, the result of this project will be able to provide key insights into the characteristics of each explanation model for further development of the AI/ML that will be utilised in the DT, which ultimately will contribute to the vision of Ecofactorij.

3.2.2 Requirements

Given the objectives and the stakeholder analysis above, the following functional and non-functional requirements have been identified. These requirements will be applied to encompass the evaluation method of WindowSHAP and TimeSHAP in time-series forecasting.

3.2.3 Functional Requirements

1. Interpretable Outputs: The output gained from the explanation models must be generated in a visual or textual form of information that is intuitively understandable to human users.
2. Model Compatibility: The prediction models need to be compatible with the explanation model, WindowSHAP and TimeSHAP
3. Ecofactorij centric data usage: The training data will be provided, and the model needs to output the intended information, which is
    - earnings on the various energy markets
    - maximum use of generated sustainable energy (solar power)
    - robustness of the grid
    - utilisation rate of batteries and the network
4. Quantitative Approach: The experiment methods need to be metric-based for the quantitative and objective evaluation
5. Human-centred Approach: The evaluation methods include a human-based experiment to test with human users to verify whether the model outputs are interpretable

3.2.4 Non-Functional Requirements

1. Explainability Clarity: The generated explanation outputs need to be comprehensible without technical knowledge of the AI/ML fields
2. Consistency: The generated explanation outputs need to stay consistent when the interaction continues with similar prediction outputs
3. Truthfulness: The generated explanation outputs must be aligned with the patterns and algorithms on which the outputs from the prediction models indeed are based

4. Feature Efficiency: The explanation methods need to interpret the patterns and algorithms of the prediction outputs with the minimum features required; in other words, the explanation methods need to precisely pick the features that are necessary for the assessment

5. Black Box Feature: Since the interest of this study lies in the assessment of the explanation methods on the black box models, the prediction models need to be opaque

3.2.5 Constraints

1. Only WindowSHAP and TimeSHAP will be evaluated, and further development will not be included

2. As of the background constraints and the research interest, which will focus on the performance of explanation methods, only the simple time-series models will be applied

3. Deployment of a practical model is not planned within the scope of this project

3.2.6 Project Goals

The overarching aim of this project is the adaptation of the TimeSHAP and WindowSHAP plot methods into a synchronised format, enabling direct and coherent comparison of their attribution outputs. Consequently, the experimental evaluation of these two methods will be conducted using distinct qualitative metrics to assess their performance across crucial aspects. This study shall conclude with the provision of a test result and the subsequent future work suggestions attained from the discussion and the analysis. In addition, the preprocessed dataset and the adapted SHAP visualisation framework used throughout the test process will be submitted to the aforementioned stakeholders as supplementary materials.

3.2.7 Development Environment

The implementation was executed on a MacBook Pro 2020 equipped with an Intel processor, 16GB of RAM, and running macOS 14.6.1. This setup ensured sufficient computational capacity for training and evaluating neural network models and SHAP explainers. The project incorporated WindowSHAP and TimeSHAP using the existing open repository published by each of the original developers [14][13]. Thereby, PyCharm served as the main development platform, with a Python 3.12 virtual environment, which is compatible with the SHAP version of both TimeSHAP and WindowSHAP and all the other required packages to train, run and evaluate the forecasting and explanation models. Regarding the data pre-processing, based on the non-disclosure agreement on the contract, all the necessary manipulation was conducted on a secure local environment using PyCharm instead of web-based computing programs represented by Jupyter Notebook. This ensured adherence to the contractual procedure, preventing the data file from being exposed to third-party entities.

3.3 Data Understanding

3.3.1 Data Description

The raw data file for this project was transferred from the preceding institution, with the contract specifying the limited use. The data file details the amount of energy supplied to each facility through the energy grid, including the supply to the office(M13) and the battery(M60), which are the main focus of this study. The raw dataset in the files consisted of individual daily files covering the period from January 1st to December 31st, 2024, in which each file contained electricity records at the office meter and battery meter. For both, distinct CSV files were aggregated based on the data collection interval, which was composed of 1m, 5m, 15m and 1h. Most of the CSV files were complete despite some minor cases, where certain rows were found to have missing numerical values.

3.4 Data Preparation

3.4.1 Net Consumption

Inside the battery and the office meter file, two important columns provide information on Apparent Instantaneous Power (obis_9_7_0) and the sum of Active Instantaneous Power (obis_16_7_0). The prior represents the compound of active power (P) that's responsible for performing useful work, and reactive power (Q) that supports voltage stability in the circuit, which could be derived by:

$$S = \sqrt{P^2 + Q^2} \tag{1}$$

The latter specifies the energy load that was consumed by the customer and the electric circuit. In order to incorporate the ML modelling, a corresponding calculation process was required in combination with these two powers to derive the actual consumption load in the grid. obis_9_7_0_mean encapsulates data on both active and reactive power, and obis_16_7_0_mean provides information reflecting the positive and nagative trends. In this project, the 15-minute interval dataset was chosen as the primary focus, as it strikes a balance between capturing short-term fluctuations accurately and preserving mid-term trend information. By combining both, a new timeseries which constitutes both pieces of information in one was derived by the following formula:

$$\text{New\_timeseries} = \text{obis\_9\_7\_mean} \times \text{sign}(\text{obis\_16\_7\_mean}) \tag{2}$$

This calculation process was implemented for both M13 and M60 meters from all daily files. Thereby, using these new time series for both battery and office, the net consumption load was

derived from the following formula:

$$\begin{aligned}\text{Net\_consumption} = {} & \text{M13\_obis\_9\_7\_mean} \times \text{sgn}(\text{M13\_obis\_16\_7\_mean}) \\ & - \text{M60\_obis\_9\_7\_mean} \times \text{sgn}(\text{M60\_obis\_16\_7\_mean})\end{aligned} \tag{3}$$

3.4.2 Data Range

Although the new dataset includes daily net consumption data for the entire year, the implementation of the full dataset was found to be impractical. It was due to the original data file containing only the numerical values specifying the energy load in the grid, excluding the potential independent variables that could affect the values. There could be numerous influencing variables, such as weather conditions, seasonal climate changes, and daylight duration. As a result, the forecasting approach was simplified to a univariate model rather than a multivariate model. However, due to the scope of this study focusing on the explainability nature of TimeSHAP and WindowSHAP, the approach taken was to narrow down the range of input to the models so that the potential independent variables can be minimised to affect. Referring to the Koninklijk Nederlands Meteorologisch Instituut (KNMI), the Dutch national meteorological institute [16], which provides yearly summaries of national weather data, a suitable period was identified for the analysis. The primary criteria included the absence of extreme weather cases, relatively stable temperature, and moderate precipitation. As a result, the selected period spans three months from March 30, 2024, to June 30, 2024.

3.5 Modelling

3.5.1 Prediction Model

The prediction models that were analysed focused on time-series models. Some of the potential candidates were conventional ML models such as ARIMA, SVM, and SARIMA[17]. However, these are not deep-learning models, and some of them are even outdated in terms of their practical usage when it comes to energy forecasting. Most importantly, non-deep learning models are rarely considered black boxes because of their characteristics that rely on statistical calculation without contextual information[18]. Therefore, Neural Network(NN) models such as Transformers, LSTM, GRU, RNN and other neural network models were deemed more relevant. Nonetheless, non-neural network models such as XGboost and Facebook Prophet have been popular among other time-series forecasting models, and they are also considered black box models. To broaden the scope of the study, both neural network-based and non-neural network models appeared worthwhile to explore. However, due to the limited comparative research concerning the performance of interpretability methods like WindowSHAP and TimeSHAP, this project ultimately focused on NN models to ensure greater effectiveness. Although CNN and other related extension models adopted for sequential data type were also considered, TimeSHAP and WindowSHAP specifically require returning the hidden state; therefore, the selected NN models must have been memory-based models. Consequently, LSTM, RNN, and GRU were selected for prediction models because of their versatility and inherent suitability for time-series tasks.

3.5.2 Model Architecture

Attribute	LSTM	GRU	RNN
Input	(samples, timesteps, features)	(samples, timesteps, features)	(samples, seq_len, 1)
Layer 1	LSTM(32)	GRU(32)	SimpleRNN(64)
Layer 2	LSTM(32)	GRU(32)	SimpleRNN(32)
Hidden Units	32	32	64, 32
Dense Layers	Dense(1)	Dense(1)	Dense(150, ReLU), Dense(1)
Output	1	1	1
Loss Function	MSE	MSE	MSE
Epochs	5	5	5
Validation	20% validation split	20% validation split	20% validation split

Table 1: Comparison of Deep Learning Models for Time-Series Forecasting

For the sample data, the aforementioned preprocessed dataset was used as its input. The data was split into an 8:2 ratio, with for the training and testing. Subsequently, for both the data parts, the labelling process was implemented to ensure that each sequence is configured to 24 timesteps in length, and the output will be derived from the one step after the sequence. To ensure fair comparison and consistency across different neural network models, a uniform architecture was applied to the LSTM, GRU, and RNN models. Each model consists of a sequential neural network structure designed specifically for time-series forecasting tasks. The architecture includes the following components: All models were trained using the Adam optimiser with a learning rate of 0.001 and mean squared error (MSE) as the loss function. To ensure compatibility with both TimeSHAP and WindowSHAP, the models were implemented using TensorFlow and PyTorch libraries for each, while maintaining identical model architectures across the two frameworks.

3.5.3 Explainability Integration

The trained prediction models with each explanation method followed integration structure as visualised in Figure 1. This architecture ensured that each model not only predicts short-term energy consumption but also provides interpretability results at the same time using the corresponding SHAP-based explainer.

3.5.3.1 TimeSHAP

The base package was installed through the original GitHub repository published by [13]. The package was a comprehensive framework that covered data splitting, model training, and SHAP plot visualisation thoroughly. To begin with, the data file, the split function, the labelling process, and the selected model architecture were modified so they aligned with the chosen methods. Figure 2 shows the plot before the visualisation modification. X-axis layouts the timepoints counted

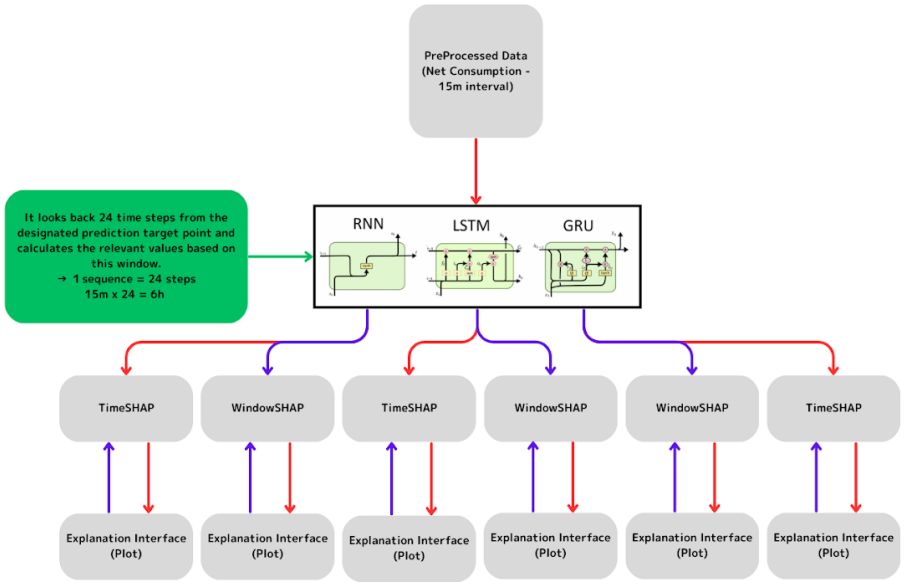

Figure 1: Prediction model and SHAP tools integration

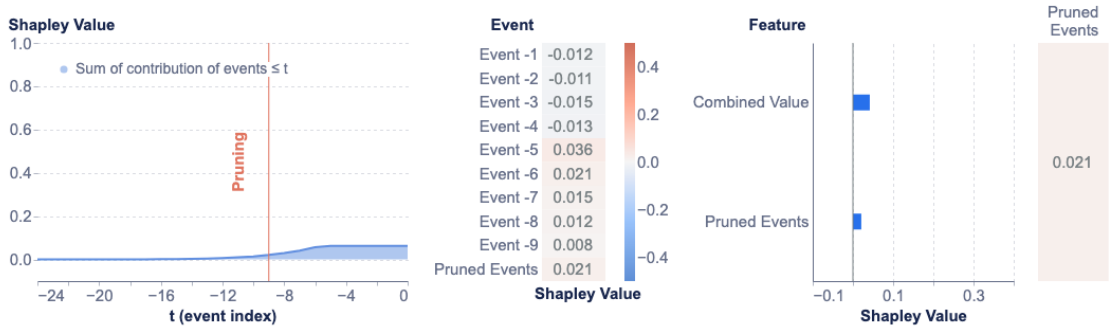

Figure 2: TimeSHAP plot

inversely from the predicted target point, which in this case is the 25th from the start of the sample input index. Y-axis summarises the sum of contributions across each timestep in the sequence, showing the gradual increase as the step proceeds. It applies the pruning function to extract the most significant part in the sequence. The middle plot in Figure 2 summarises the contribution of each timestep after the pruning process, making it clear and intuitive to grasp the significance of the individual timesteps. The last plot on the right side of Figure 2 summarises the contribution of features, which in this case has only two, as the forecasting model is univariate. In its current form, it presented difficulties in comprehending the plot information simultaneously and intuitively understanding how much exactly each timestep weighed against the observed value from the sample. Accordingly, the visualisation was modified to integrate the content of the middle and left plots into a single graph for easier interpretation. The modification process involved using the same event data originally employed in the middle plot of Figure 2, transforming it into individually colored dots, each colour encoding the magnitude of the distributed SHAP values. The goal was to synthesise these dots with a line chart representing the absolute observed values from the sam-

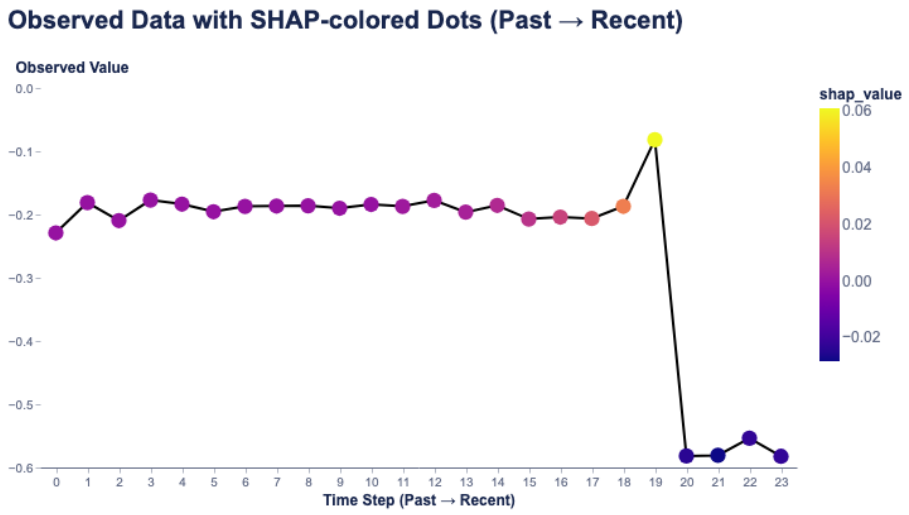

Figure 3: TimeSHAP plot after modification

ple input, all within a single, unified frame. As a result, Figure 3 shows the modified plot where the observed value is deployed, and on top of each bending point of the graph, the dots configure the SHAP magnitude by heat map colour. Additionally, the time sequence was adjusted so that the time step proceeds forward starting from zero on the x-axis.

3.5.3.2 WindowSHAP

As it was done for TimeSHAP, the original code package was referred to the author [14] and

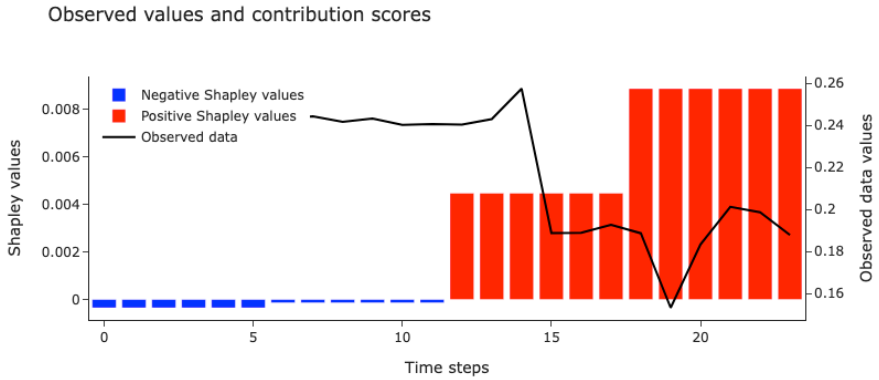

Figure 4: WindowSHAP plot

installed following the manual through the public GitHub repository. Based on the identical procedure as TimeSHAP development, the defined model architecture and the data processing were implemented. Figure 4 is a sliding window plot, one of the WindowSHAP functions that generate SHAP values by shifting each window per timestep. In this way, it prevents WindowsHAP from overlooking the effect of transition between each partition [14]. Although it already resembled the modified visualisation plot done for TimeSHAP, for a fair evaluation, the visualisation style had to be uniform. For WindowSHAP, the observed value was already placed in the form of a line chart, and the corresponding SHAP contributions were overlaid in the same frame. Thereby, the

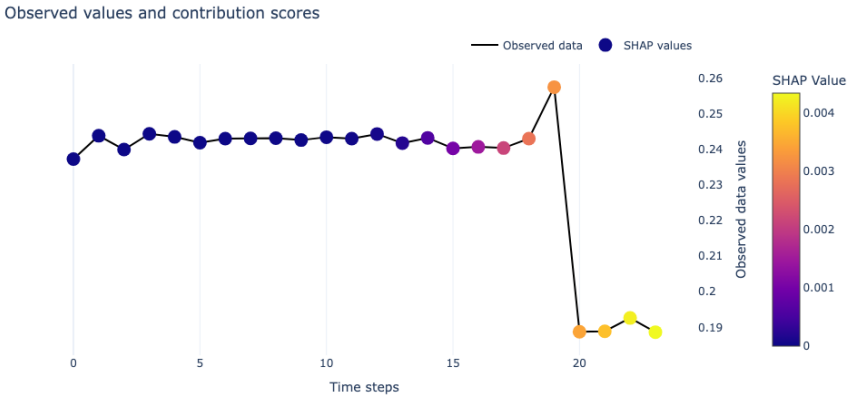

Figure 5: WindowSHAP plot after modification

modification process involved changing the plot style of SHAP from bar chart to dot type while ensuring that each dot comes on top of the bending points of the line chart. Consequently, Figure 5 shows the modification result. It reflects the coherent visualisation format shared with Figure 3. Although both use the identical heat map spectrum, a slight difference can be found in the dot colours. It was due to TimeSHAP and WindowSHAP incorporating different plot libraries from each other. For the dot location, it was accomplished by synchronising the y values with those of the observed values, which shared the same point on the x-axis.

3.6 Evaluation

3.6.1 Evaluation Objective

The overarching aim of the evaluation was to assess the two distinct variants of SHAP—WindowSHAP and TimeSHAP—based on their merits and demerits across their characteristics. Although both of the methods share the foundational concept in their design, they are suitable for sequential data processing; the evaluation attempted to identify their gap in capability.

3.6.2 Metrics

The evaluation methodology involved quantitative metrics such as fidelity, consistency, and stability. Fidelity was a crucial factor in assessing the explanation method because it measures how robustly and precisely the interpretations adhered to the decision-making process of the prediction model in a sensical manner [19]. Likewise, sparsity plays an important role in the comparison of the performance, revealing the models' ability to accurately describe the model output using the minimum amount of yet significant features [20]. Lastly, stability is another impactful indicator measuring how truthful the explanation models are by confirming how often the generated explanations stayed consistent and are not affected by trivial modifications of the input from the sample [21]. In other words, stability refers to the likelihood of an explanation model returning a consistent output from a similar dataset.

3.6.3 Evaluation methods

3.6.3.1 The Likert Scale

To conduct the evaluation process, this study employed the Likert scale survey as the primary method of data collection. The Likert scale survey is one of the prominent psychometric tools in the social and educational research domain. It was primarily designed to quantify subjective measures like perceptions, which are often difficult to estimate [22]. The typical Likert scale enables such quantification by having the participants answer survey questions on a 5 to 10 scale, where the lowest value represents their strong disagreement and the highest value represents their strong agreement with the questions. In the form, the choices were coded as follows:

- 1 = Strongly Disagree
- 2 = Disagree
- 3 = Neutral
- 4 = Agree
- 5 = Strongly Agree

3.6.4 Metrics conversion

To support this evaluation, the questions were generated based on the preceding three metrics. These metric-based questions were designed to ensure that the participants understand the fundamental concepts underlying the evaluation criteria without needing to know the specific definitions of each metric term. Additionally, sub-questions concerning the overall intuitiveness and cognitive load of the modified visualisation were appended aside from the key metric inquiries.

3.6.5 Participants

The evaluation involved 3 participants from the ML, AI-related academic and professional domains with foundational background knowledge of and experiences with SHAP methods. One of the participants held previous experience with the TimeSHAP and WindowSHAP methods prior to this evaluation.

Prior to commencing the evaluation, the invitation forms were sent to each participant via email, and all of them consented to be involved in this evaluation. The survey adhered to the standard guidelines of GDPR [23], thereby no personal information and data were collected during the process.

3.6.6 Procedure

To support the evaluation, the example plots were generated using TimeSHAP and WindowSHAP with the aforementioned visualisation techniques. For every combination of the three prediction models and the two explanation methods, the graphs were generated based on five distinct target indices: 0, 20, 50, 500, and 1000, allowing participants to grasp the tendencies of the explanation

models by providing their movement at multiple time points, and prepare them to respond to the survey questions. In total, 30 example plots were created.

After the preparation of the materials mentioned above, an invitation form was distributed to the participants. The form outlined the purpose of the study and contained a request for informed consent. Upon receiving confirmation from the participants, participants were provided with the link to an online form containing the survey questions as well as a cumulative file of all the example plots for reference purposes and the introductory manual specifying the conduct guidelines to proceed further.

Once the link had been received, the participants were allowed to start the evaluation session guided by the attached manual. The participants were asked to review the example plots in the file prior to proceeding to the survey questions. After reviewing them, they were asked to fill in the survey question with choices on a five-point scale. After the participants had finished reviewing all the questions, they were kindly asked to return the answers by closing the form.

3.6.7 Assessment of Scores

After gathering all the participants' responses, the survey data were analysed and summarised by using descriptive statistics appropriate for ordinal-scale data. Rather than calculating means or composite scores, the analysis focused on ascertaining the median and mode of participant responses per metric question, as means cannot be directly used to measure the central tendencies [24]. This approach would have provided a more accurate portrait of central tendencies in Likert-scale data; however, the number of participants was limited in this attempt, and the effect of this was restricted. Consequently, the spread of the responses into different categories (e.g., agreement vs. disagreement) was visualised using bar charts to assess the results. The interpretation of scores was thus addressed by direct visual inspection and descriptive comparison instead of statistical generalisation.

3.6.8 Constraints

One of the constraints that emerged during the evaluation was the scarcity of participants due to the nature of this assessment requiring experts with prior knowledge in the xAI domain as its respondents. In this study, only three participants were invited; thereby, this could affect the generalisability of the obtained results. Furthermore, the Likert scale survey depends on the participants' subjective perception; therefore, their prior knowledge and experience with SHAP methods may have affected the result.

3.7 Deployment

Within the scope of this project, practical deployment of any means will not be involved, as the objective of this conduct and the result will be limited to the realm of experimental attempt. However, the knowledge and techniques acquired from this study will be discussed and evaluated, and future implementation will be suggested at the end of the study.

4 Results

4.1 General Outline

TimeSHAP outperformed WindowSHAP with higher marks for all three evaluation metrics of the methods. The result implies the overall superiority of the TimeSHAP utility in univariate energy load forecasting, although the score differences were marginal.

4.2 Evaluation Results

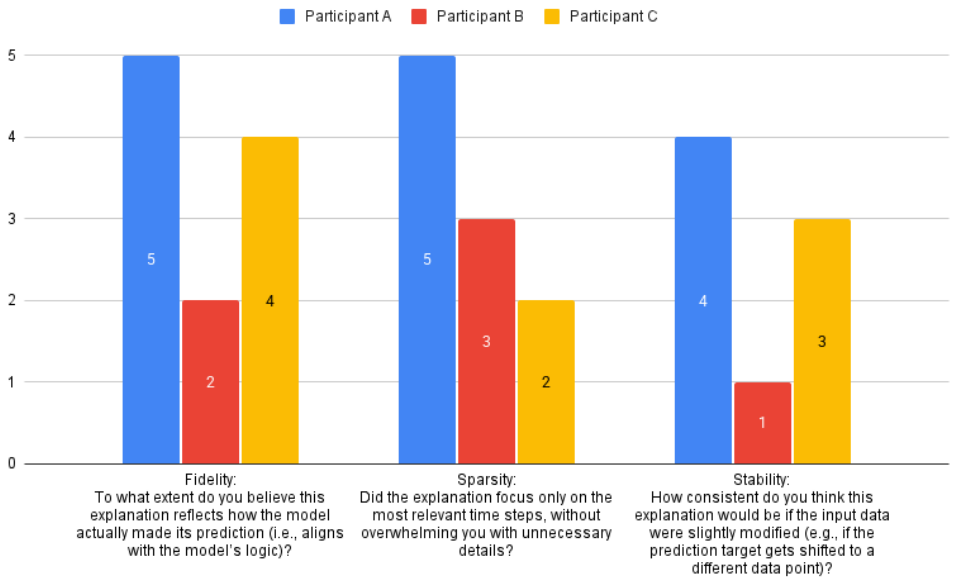

Figure 6: Metric survey result - TimeSHAP

This section presents the findings derived from the Likert-scale survey responses. The results record participants' opinions on the two explanation methods, TimeSHAP and WindowSHAP, against the above metrics, along with overall intuitiveness and cognitive load. The evidence is summarised using descriptive statistics such as median and mode, and supported with visualisations showing the distribution of responses along each evaluation metric. While the small sample size of participants precludes formal statistical inference, the results can provide preliminary qualitative indications of the relative performance and interpretability of explanation methods under consideration.

Figure 6 and Figure 7 show the visualisation of TimeSHAP and WindowSHAP survey results converted into a bar chart. The question script asked the respondents to indicate on the X-axis, and the respective metrics used are indicated. The highlighted points are plotted on the y-axis on a five-point scale.

For the Fidelity metric, the highest score was marked 5, and the score of 2 was given as the lowest among the others, leaving the median of 4. Meanwhile, WindowSHAP reflected a slightly diverging result of 3 being the highest and 2 being the lowest, with the median of 3. The notable

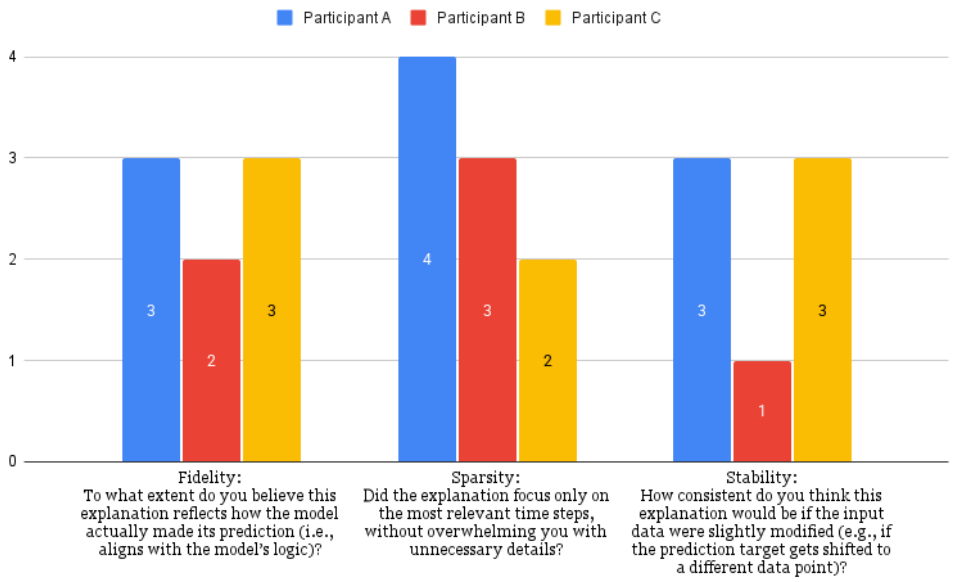

Figure 7: Metric survey result - WindowSHAP

point here is the appearance of participants A and C, who marked higher for the TimeSHAP, but both marked one point lower for the WindowSHAP. Participant B shared similar views on the Fidelity metric for both methods. This suggests that Individual preferences for model fidelity might differ depending on the explanation method. Nonetheless, some users might consider TimeSHAP to be a bit more consistent or interpretable.

In terms of the Sparsity metric, only Participant A noticed a difference, rating TimeSHAP one point higher than WindowSHAP. This indicates a lack of comparability. The small difference suggests that most participants felt that both techniques were roughly equivalent in the degree of sparseness or selectivity of the features that were shown. Still, Participant A's rating implies that TimeSHAP did a better job of focusing on key concepts while ignoring irrelevant material. This could be due to TimeSHAP's focus on temporal contributions, which results in clearer, sparser encodings. In contrast, WindowSHAP might mask effective contributions through its window-based aggregation.

On the stability measure, only Participant A reported a difference in stability. They scored TimeSHAP slightly better than WindowSHAP. It is interesting because this is also the participant who preferred TimeSHAP for the Sparsity metric. It might indicate some idiosyncratic bias or the view that TimeSHAP's results are less volatile with respect to changes in the input. More notably, Participant B gave both methods the lowest possible score of 1, which indicates a strong perception of instability. This implies that, independent of the approach, the explanations might have appeared to be inconsistent or overly sensitive to deviations in model behaviour or data conditions. A low mark may reflect a limitation in the way that TimeSHAP and WindowSHAP explanations are conceived or delivered, at least in the view of that participant. Participant C rated both approaches moderately, reinforcing the overall impression that neither approach clearly outperformed the other on perceived stability. While responses such as Participant A's slightly

favour TimeSHAP, the overall trend is that stability is an issue for both explanation approaches.

4.3 Sub-question Results

In addition to the main metric questions above, the following questions were asked about the overall preference and preliminary impressions of both methods at the end of the survey. The asked questions are summarised as follows:

- **Q1**: Was the visualisation format (e.g., dot plot application, layout) helpful for understanding the importance of time steps?
- **Q2**: How mentally demanding was it to interpret this visualisation?
- **Q3**: Which method (TimeSHAP or WindowSHAP) helped you better understand the model's behaviour from a broader, time-series perspective?
- **Q4**: Conversely, which method helped you better understand the reason behind a specific decision at a particular time step?

For both of the explanation methods on a five-point scale, one of the participants found the dot application of the plot employed in this study helpful, while the others disagreed. One point of reflection from this would be that during the evaluation process, the unmodified plot was not available to the participants; therefore, it might have been difficult for them to distinguish the general competencies. This aspect of the evaluation could be expanded and reassessed by a specific comparative analysis on the visualisation format, focusing on its effectiveness in the further implementation.

For the second inquiry, all of the participants found the adopted visualisation cognitively challenging. This may reflect the constraints in resource availability and the limited contextual background provided for the sample data and its intended forecasting use in this attempt.

For the last two questions, the overall subjective impression of the two methods was inquired into. While the question result showed a little margin of difference in the model's capability in providing insights from a broader perspective, two of the participants exhibited a preference for TimeSHAP over its explainability at specific time steps.

4.4 Limitation

Although this evaluation showed the effectiveness of TimeSHAP based on the proposed metrics, there remain several limitations throughout the evaluation process. As mentioned in the previous chapter, the number of participants was restricted; therefore, the generalizability of the results is questionable. Furthermore, based on the confidential agreement, the background information underlying these forecasting models and the processed data were limited to participants, which may have influenced their overall understanding of each plot. After the evaluation, one of the participants provided feedback expressing their struggle throughout the process for this reason.

5 Future Work

This study explored time-step specific Adaptation of the plot visualisation for TimeSHAP and WindowSHAP, and the consecutive evaluation of the two methods by incorporating them. It attempted to reveal the explainability of TimeSHAP and WindowSHAP methods from three different distinct metric perspectives. Further analysis on these two methods may include multivariate prediction models, which this study exclusively focused on univariate models. When applied to multivariate forecasting, especially the sparsity nature of the methods holds potential to exhibit alternative results due to the greater number of features. Thereby, both TimeSHAP and WindowSHAP will be required to reflect accurately not only how much at each timestep but also which of the features. Accordingly, the timestep-based visualisation format that was achieved during this study can be enhanced to project feature information and its individual significance in magnitude.

While this study employed some of the basic memory-based NN models, the other types of extended prediction models, such as Bidirectional LSTM, can be applied and experimented with. Especially, it would be interesting to see the performance of these methods by applying prediction models that are capable of retaining a greater span of memory. In addition, the data span of the input sample can be extended to reveal more characteristics that may have been missed in this study due to constraints. While this study investigated the effectiveness of TimeSHAP and WidonwSHAP on the observed energy load dataset, another future work can study these methods with an alternative sequential dataset.

Since this analysis solely relied upon the subjective perception of the methods in the evaluation, a more quantitative/statistical approach can be applied in future work to explore the margins of TimeSHAP and WindowSHAP in greater detail. In particular, the comparative studies between these two methods remain limited in number.

6 Conclusion

This research aimed to compare the explainability and effectiveness of TimeSHAP and WindowSHAP AI techniques in short-term energy forecasting using time-series predictive models. The study used recurrent neural networks (RNNs, LSTMs, and GRUs) to practically evaluate these techniques and assess how easy their explanations are to understand.

The evaluation included quantitative measures such as fidelity, sparsity, and stability, along with a Likert-scale survey to capture subjective views on interpretability and cognitive load related to the transformation and visualisation outputs. Results showed that TimeSHAP performed slightly better across all three measures, especially in fidelity, suggesting it provides more accurate and consistent explanations for univariate forecasting. However, the differences between the two methods were generally small.

Participants' feedback revealed that both visualisation methods involved a cognitive workload, with no clear agreement on which was more intuitive. This indicates that SHAP-based explanations are challenging to understand and that the evaluation had limited resources and user support. Notably, most participants found TimeSHAP more useful for explaining model decisions at spe-

cific time steps, while neither method clearly stood out in helping users understand behaviour over longer periods.

Finally, this project contributes to the growing discussion on explainable AI for energy forecasting by empirically comparing two promising methods and offering practical improvements for visualisation. The findings lay the groundwork for future research that could include multivariate models, more complex memory-based architectures, and larger user studies to enhance the explainability and usability of AI decision support systems.

Use of AI

During the process of this study, the author accessed ChatGPT to assist with grammatical/formatting error rectification on this paper. The author inspected any modifications made by these tools and corrected the content respectively; therefore, the author bears full responsibility for the subject matter.

References

[1] Ecofactorij, "Sustainable energy management in factories," 2024, accessed: 22-Mar-2025. [Online]. Available: https://ecofactorij.nl/

[2] Tableau Software, "Time series forecasting: Predicting the future with Time-Series data," https://www.tableau.com/analytics/time-series-forecasting, 2025, accessed: 2025-04-16.

[3] B. Brożek, M. Furman, M. Jakubiec, and B. Kucharzyk, "The black box problem revisited: Real and imaginary challenges for automated legal decision making," *Artificial Intelligence and Law*, vol. 32, pp. 427–440, 2024. [Online]. Available: https://doi.org/10.1007/s10506-023-09356-9

[4] D. Mane, A. Magar, O. Khode, S. Koli, K. Bhat, and P. Korade, "Unlocking machine learning model decisions: A comparative analysis of lime and shap for enhanced interpretability," *Journal of Electrical Systems*, vol. 20, no. 2s, pp. 598–613, 2024. [Online]. Available: https://journal.esrgroups.org/jes/article/view/1768

[5] A. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, G. Menegaz, and K. Lekadir, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, Jan. 2025, arXiv:2305.02012 [stat]. [Online]. Available: http://arxiv.org/abs/2305.02012

[6] Y. Zhang, R. Ma, J. Liu, X. Liu, O. Petrosian, and K. Krinkin, "Comparison and Explanation of Forecasting Algorithms for Energy Time Series," *Mathematics*, vol. 9, no. 21, p. 2794, Nov. 2021. [Online]. Available: https://www.mdpi.com/2227-7390/9/21/2794

[7] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal Data Meets LLM – Explainable Financial Time Series Forecasting," Jun. 2023, arXiv:2306.11025 [cs]. [Online]. Available: http://arxiv.org/abs/2306.11025

[8] S. Ahmed, M. S. Kaiser, M. Shahadat Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions," *IEEE Access*, vol. 13, pp. 37 370–37 388, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10583856/

[9] D. Garreau and U. v. Luxburg, "Explaining the Explainer: A First Theoretical Analysis of LIME," Jan. 2020, arXiv:2001.03447 [cs]. [Online]. Available: http://arxiv.org/abs/2001.03447

[10] B. Raufi, C. Finnegan, and L. Longo, "A Comparative Analysis of SHAP, LIME, ANCHORS, and DICE for Interpreting a Dense Neural Network in Credit Card Fraud Detection," in *Explainable Artificial Intelligence*, L. Longo, S. Lapuschkin, and C. Seifert, Eds. Cham: Springer Nature Switzerland, 2024, vol. 2156, pp. 365–383, series Title: Communications in Computer and Information Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-63803-9_20

[11] S. Mitra and L. Gilpin, "The XAISuite framework and the implications of explanatory system dissonance," Apr. 2023, arXiv:2304.08499 [cs]. [Online]. Available: http://arxiv.org/abs/2304.08499

[12] R. Saluja, A. Malhi, S. Knapič, K. Främling, and C. Cavdar, "Towards a Rigorous Evaluation of Explainability for Multivariate Time Series," Apr. 2021, arXiv:2104.04075 [cs]. [Online]. Available: http://arxiv.org/abs/2104.04075

[13] J. Bento, P. Saleiro, A. F. Cruz, M. A. T. Figueiredo, and P. Bizarro, "TimeSHAP: Explaining Recurrent Models through Sequence Perturbations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug. 2021, pp. 2565–2573, arXiv:2012.00073 [cs]. [Online]. Available: http://arxiv.org/abs/2012.00073

[14] A. Nayebi, S. Tipirneni, C. K. Reddy, B. Foreman, and V. Subbian, "WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values," *Journal of Biomedical Informatics*, vol. 144, p. 104438, Aug. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046423001594

[15] Data Science PM, "CRISP-DM for Data Science," https://www.datascience-pm.com/wp-content/uploads/2024/12/CRISP-DM-for-Data-Science-2025.pdf, 2025, accessed: 2025-04-17.

[16] Royal Netherlands Meteorological Institute (KNMI), "Maand- en seizoensoverzichten," https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/, 2025, accessed: Jul. 15, 2025.

[17] A. R. S. Parmezan, V. M. Souza, and G. E. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Information Sciences*, vol. 484, pp. 302–337, May 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0020025519300945

[18] A. Trivedi, "Explainable forecasting models for load forecasting," Master's thesis, University of Twente, Enschede, Netherlands, 2024. [Online]. Available: https://essay.utwente.nl/100847/

[19] M. Miró-Nicolau, A. Jaume-i Capó, and G. Moyà-Alcover, "A comprehensive study on fidelity metrics for XAI," Jan. 2024, arXiv:2401.10640 [cs]. [Online]. Available: http://arxiv.org/abs/2401.10640

[20] T. Funke, M. Khosla, M. Rathee, and A. Anand, "Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks," Mar. 2022, arXiv:2105.08621 [cs]. [Online]. Available: http://arxiv.org/abs/2105.08621

[21] J. Ribeiro, L. Cardoso, V. Santos, E. Carvalho, N. Carneiro, and R. Alves, "How Reliable and Stable are Explanations of XAI Methods?" Jul. 2024, arXiv:2407.03108 [cs]. [Online]. Available: http://arxiv.org/abs/2407.03108

[22] A. Joshi, S. Kale, S. Chandel, and D. Pal, "Likert Scale: Explored and Explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, pp. 396–403, Jan. 2015, publisher: Sciencedomain International. [Online]. Available: https://journalcjast.com/index.php/CJAST/article/view/381

[23] EUR-Lex, "General data protection regulation (gdpr) – official legal text," https://gdpr-info.eu/, 2024, accessed: Jul. 16, 2025.

[24] University of St Andrews, "Statistics support: Likert scale analysis," https://www.st-andrews.ac.uk/ceed/study-skills/mathsandstatisticssupport/statisticssupport/, 2025, accessed: Jul. 16, 2025.

Appendix A

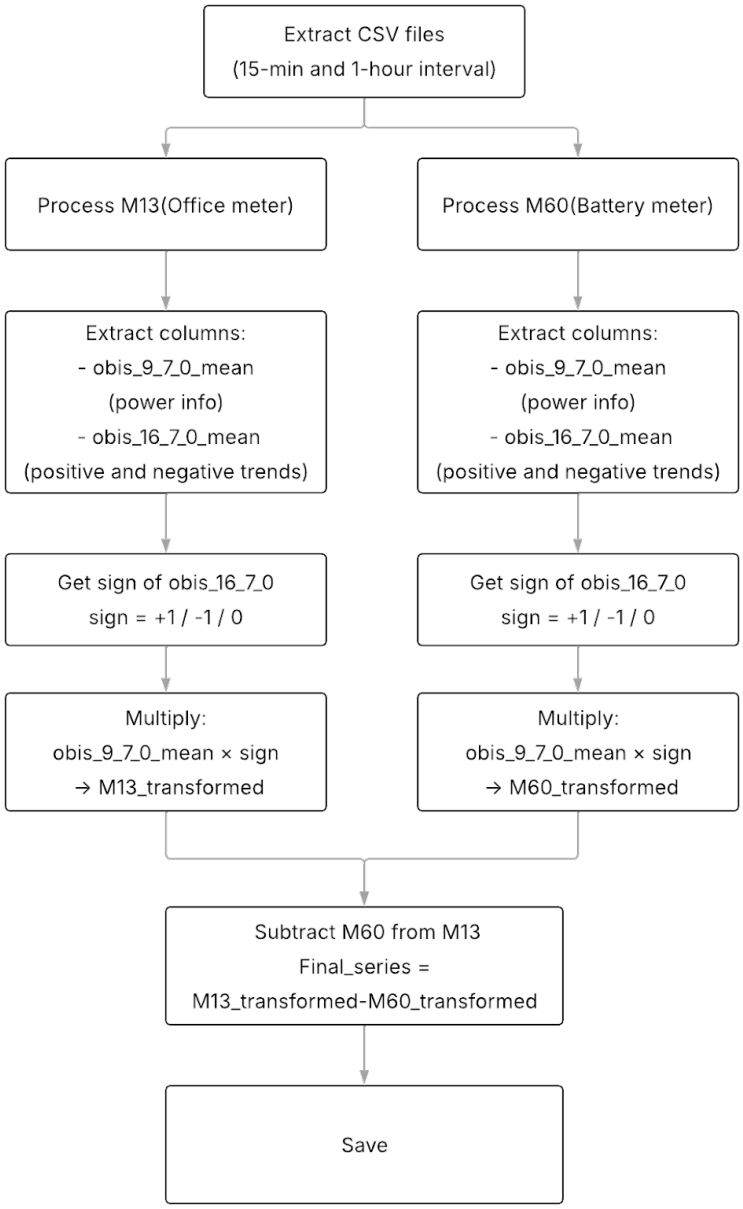

Figure 8: Flowchart of net consumption load calculation

Appendix B

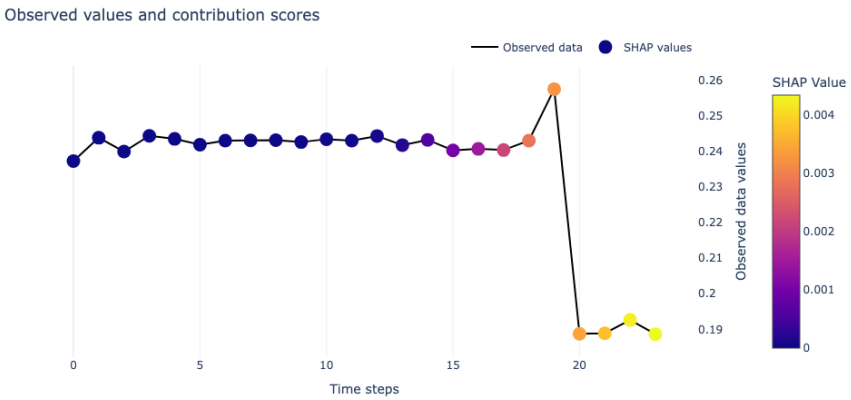

Figure 9: WindowSHAP - GRU at test index 0

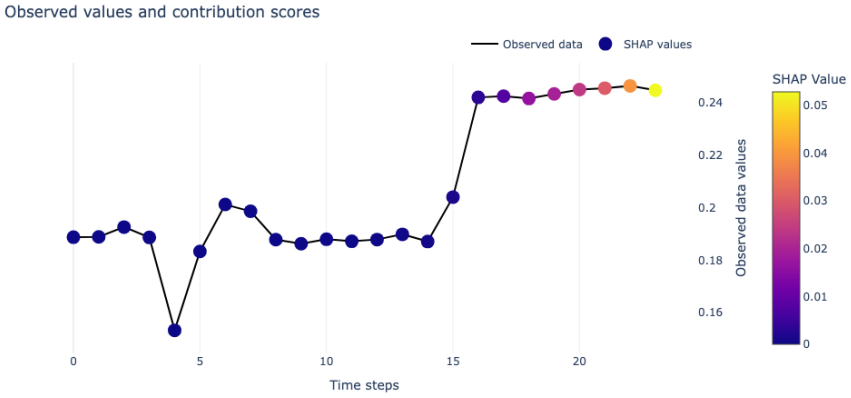

Figure 10: WindowSHAP - GRU at test index 20

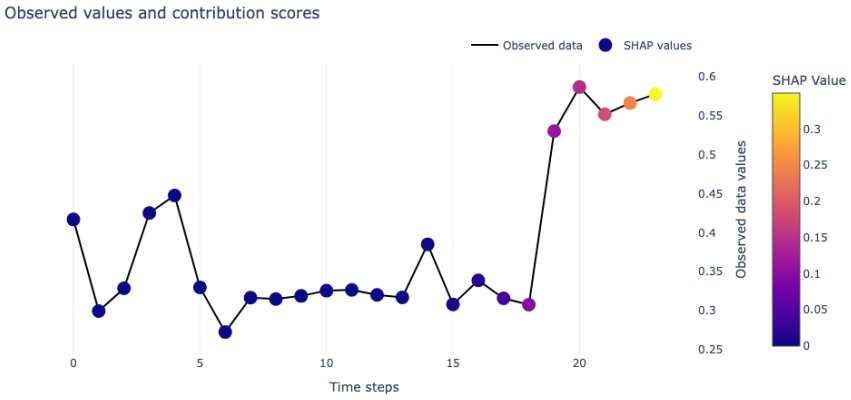

Figure 11: WindowSHAP - GRU at test index 50

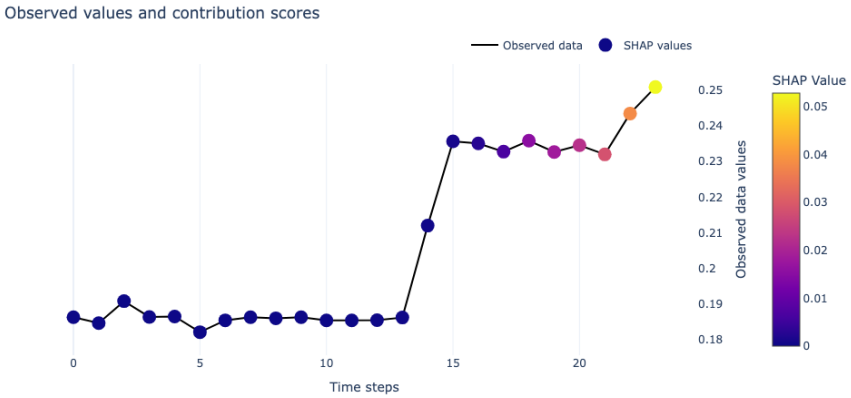

Figure 12: WindowSHAP - GRU at test index 500

Figure 13: WindowSHAP - GRU at test index 1000

Figure 14: WindowSHAP - LSTM at test index 0

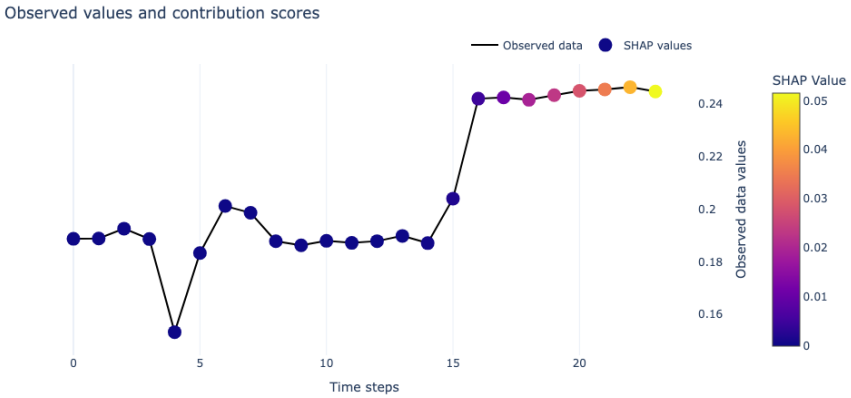

Figure 15: WindowSHAP - LSTM at test index 20

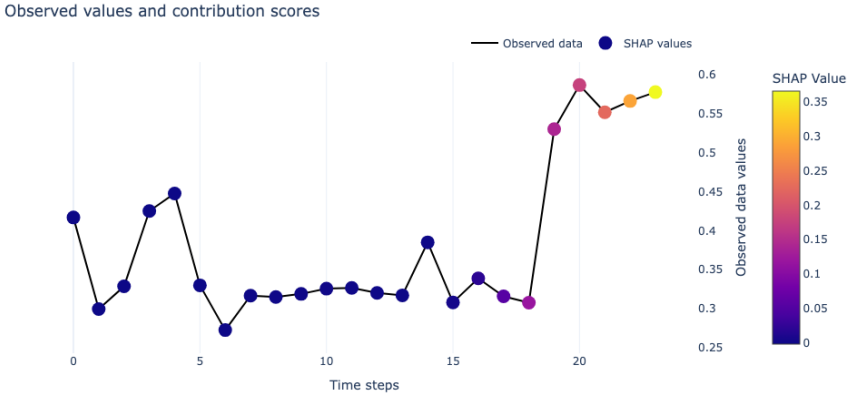

Figure 16: WindowSHAP - LSTM at test index 50

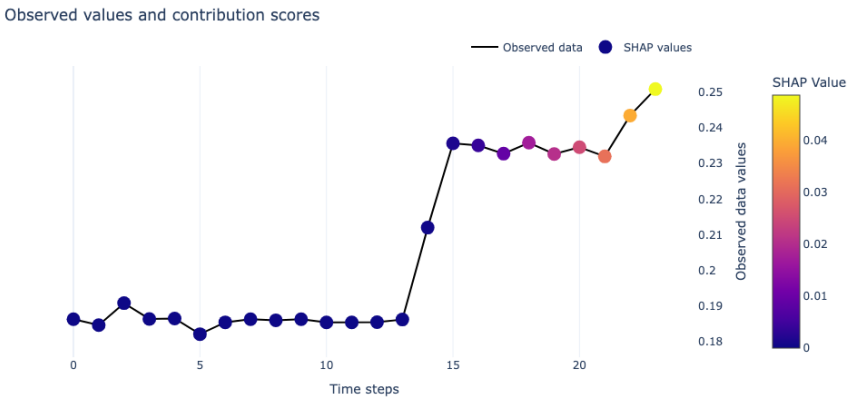

Figure 17: WindowSHAP - LSTM at test index 500

Figure 18: WindowSHAP - LSTM at test index 1000

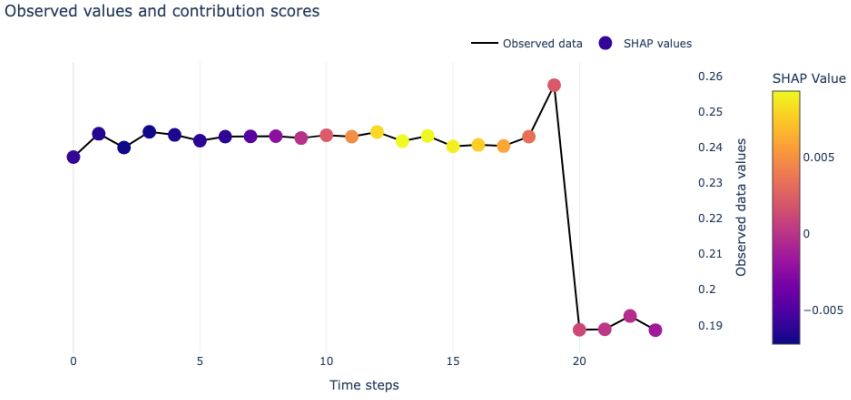

Figure 19: WindowSHAP - RNN at test index 0

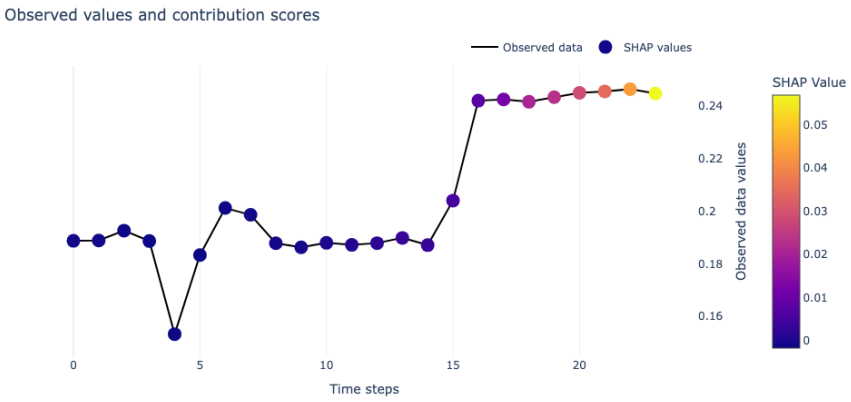

Figure 20: WindowSHAP - RNN at test index 20

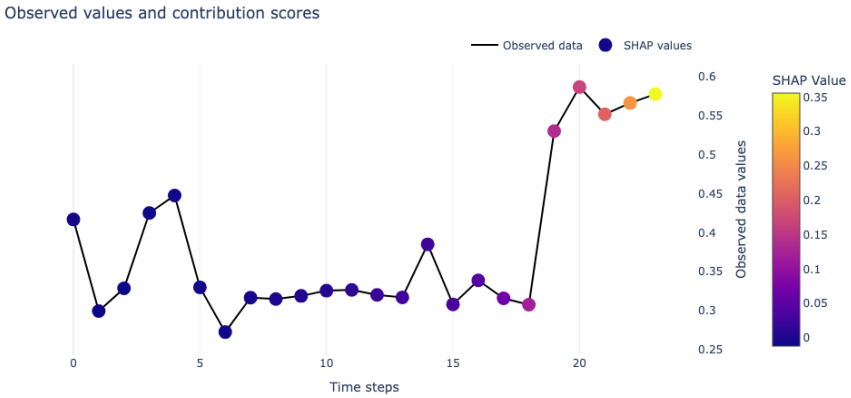

Figure 21: WindowSHAP - RNN at test index 50

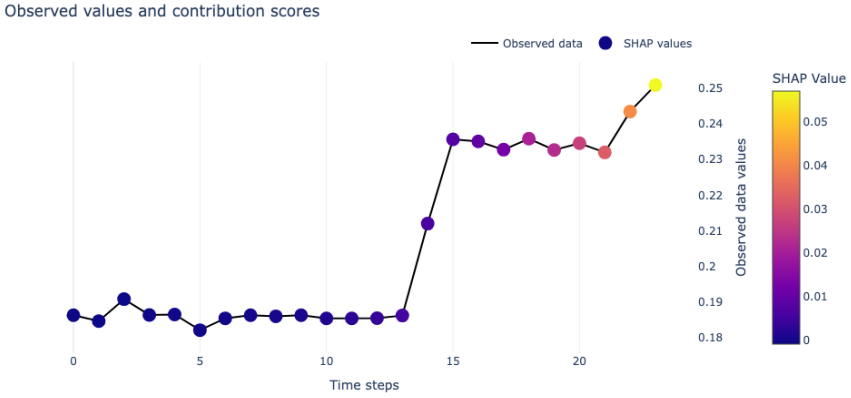

Figure 22: WindowSHAP - RNN at test index 500

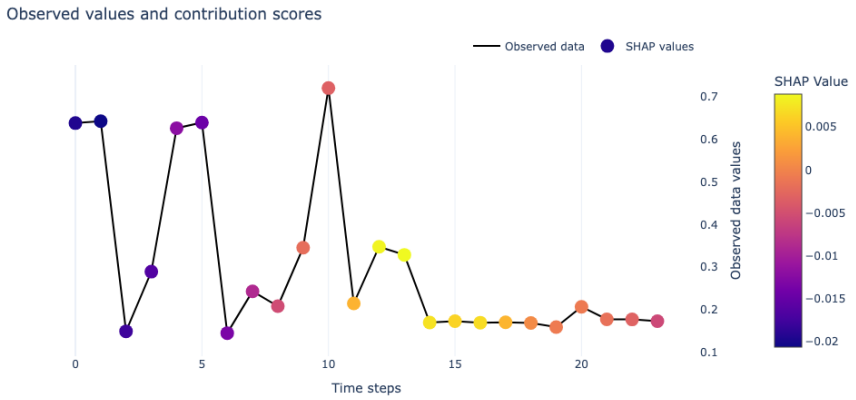

Figure 23: WindowSHAP - RNN at test index 1000

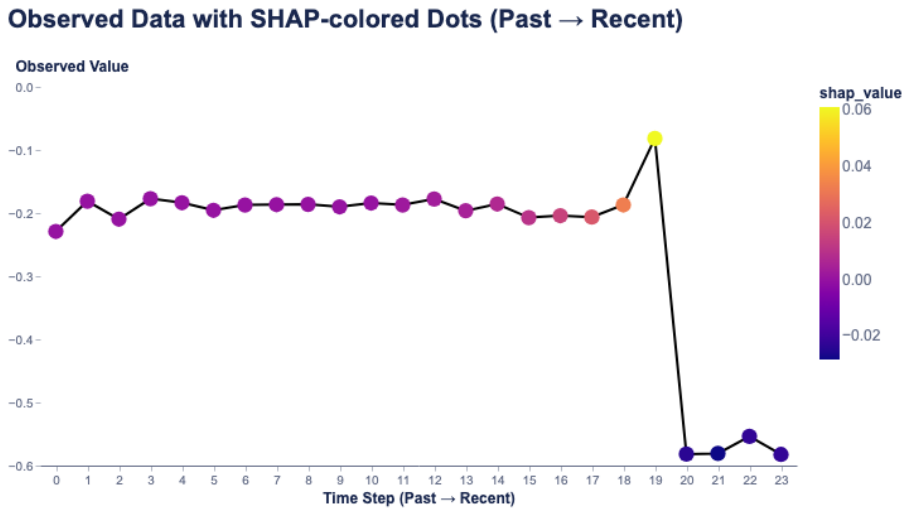

Figure 24: TimeSHAP - GRU at test index 0

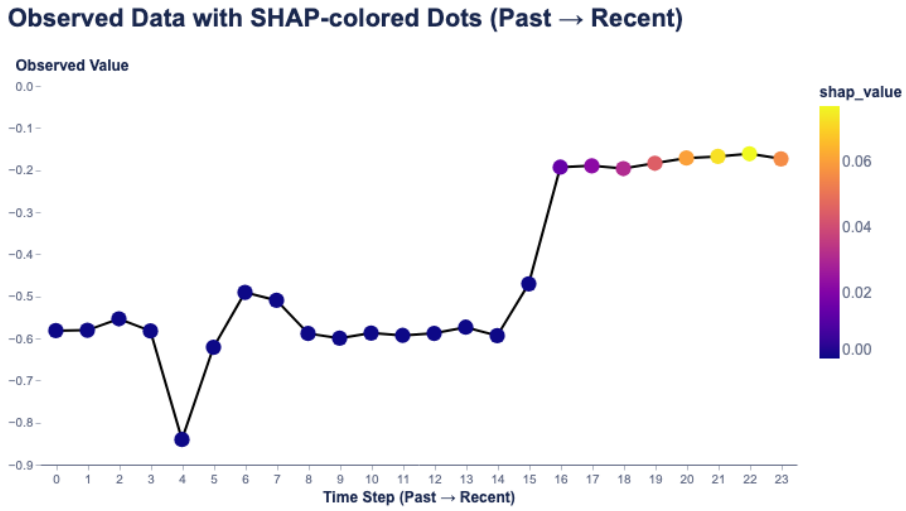

Figure 25: TimeSHAP - GRU at test index 20

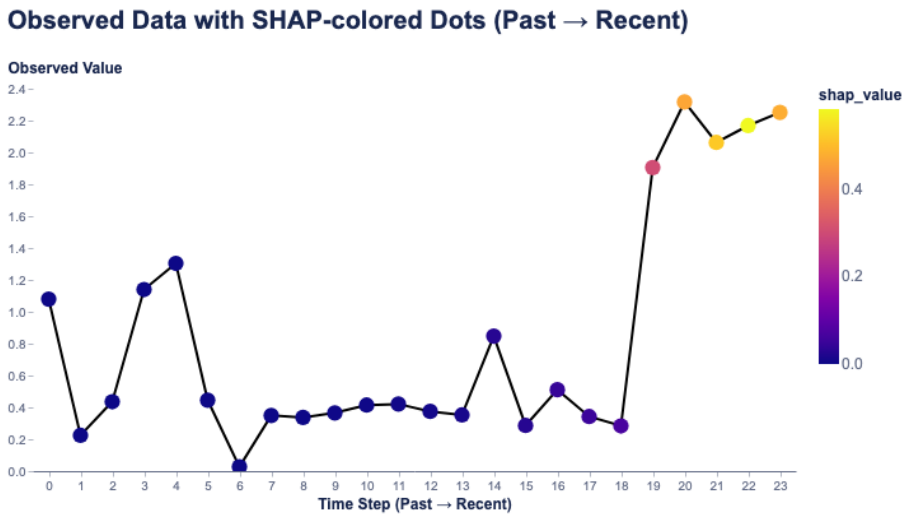

Figure 26: TimeSHAP - GRU at test index 50

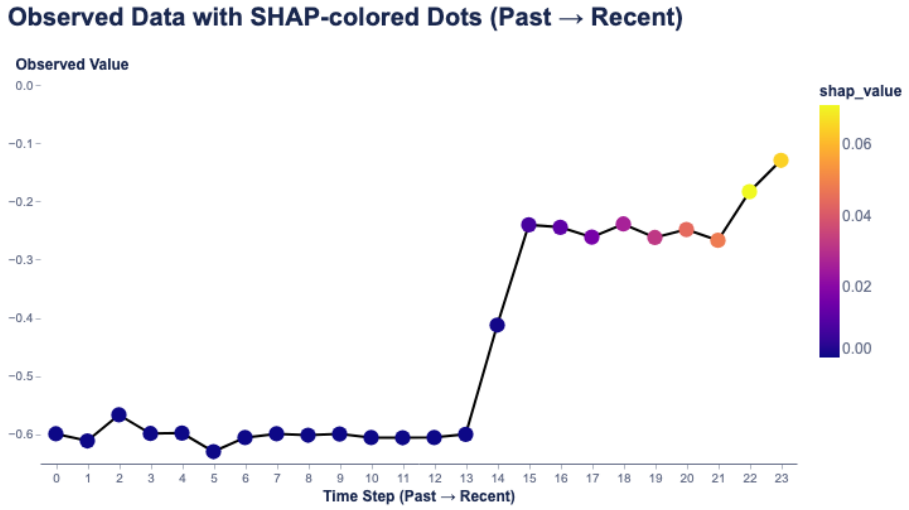

Figure 27: TimeSHAP - GRU at test index 500

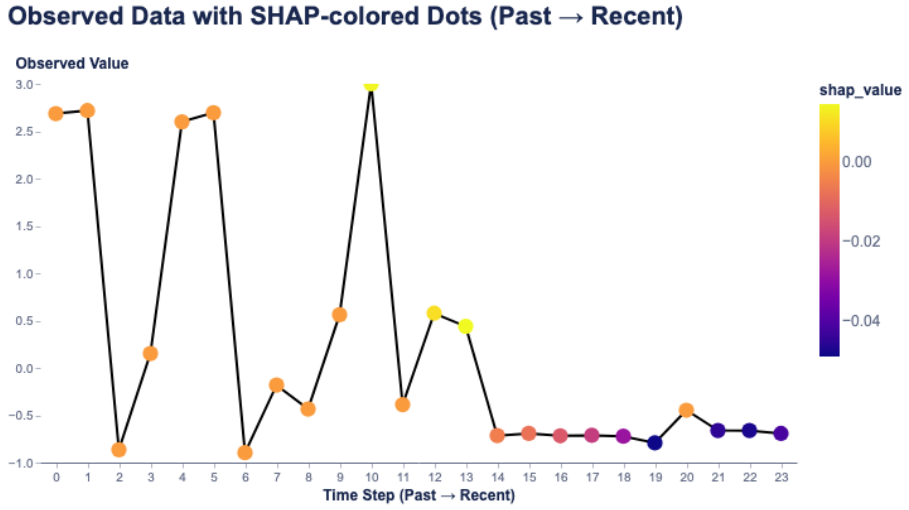

Figure 28: TimeSHAP - GRU at test index 1000

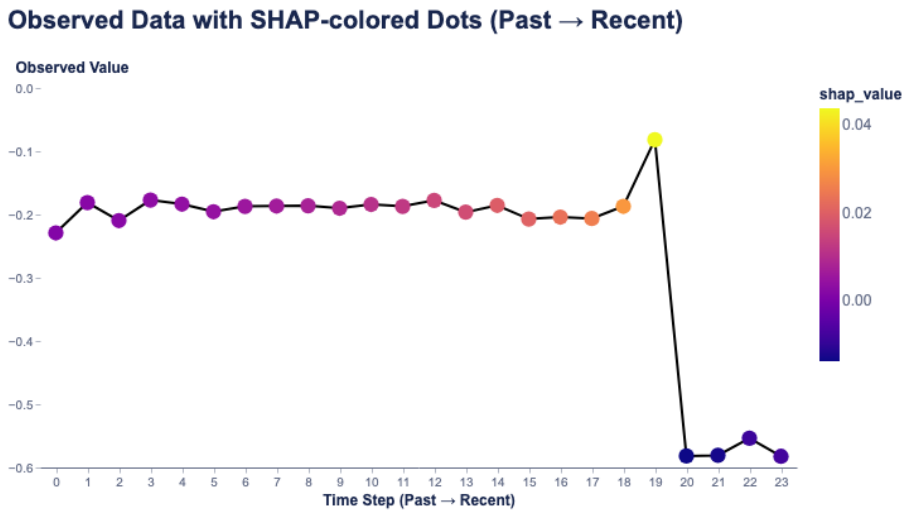

Figure 29: TimeSHAP - LSTM at test index 0

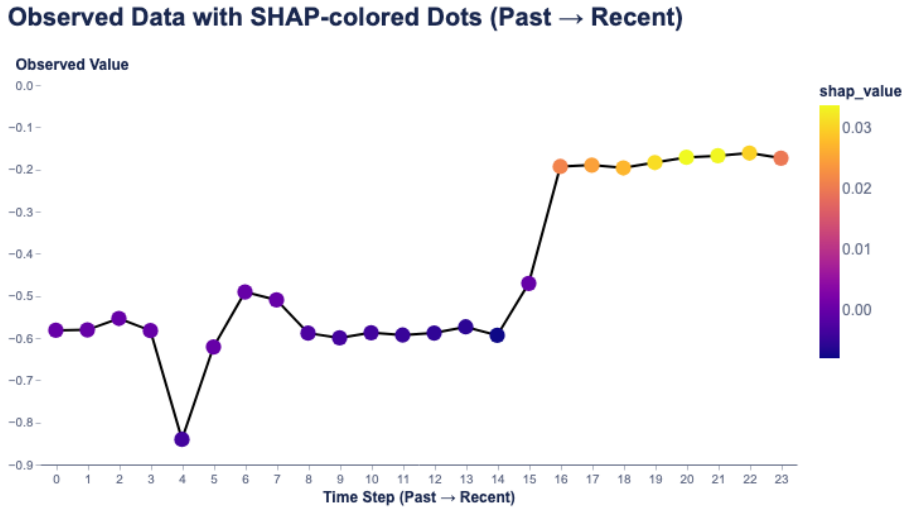

Figure 30: TimeSHAP - LSTM at test index 20

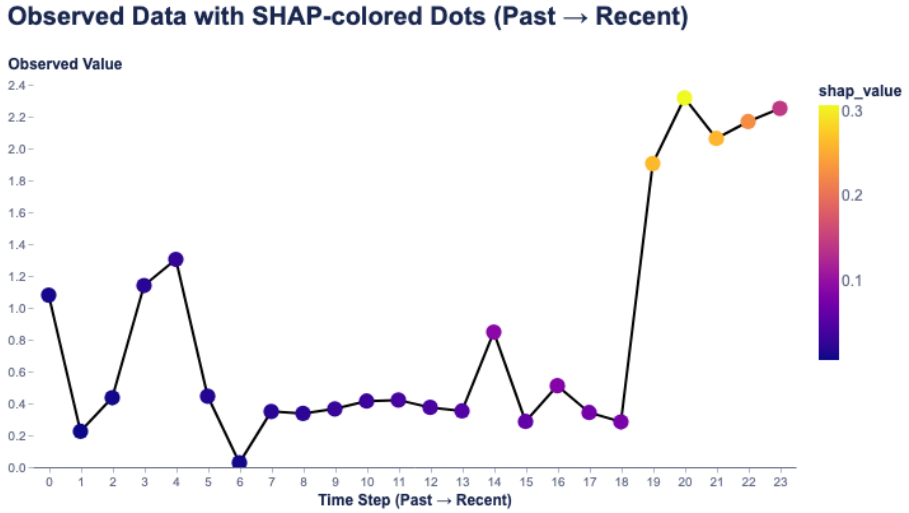

Figure 31: TimeSHAP - LSTM at test index 50

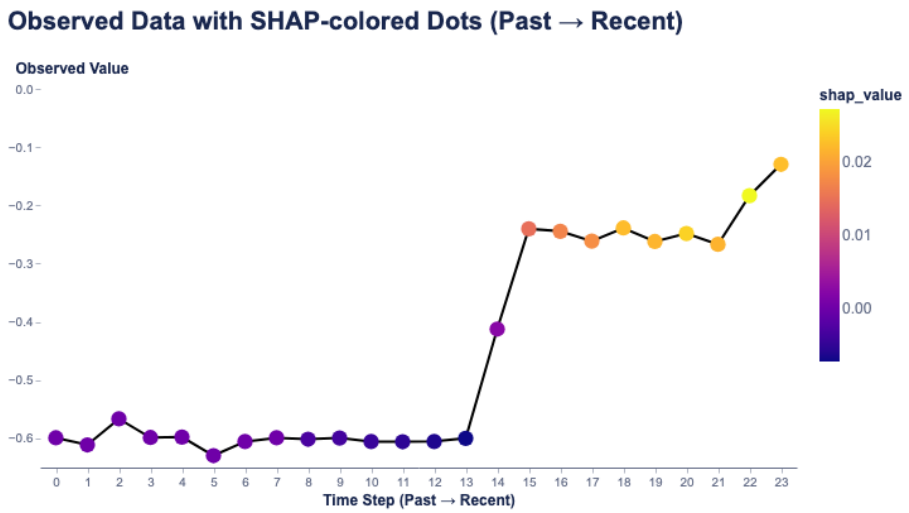

Figure 32: TimeSHAP - LSTM at test index 500

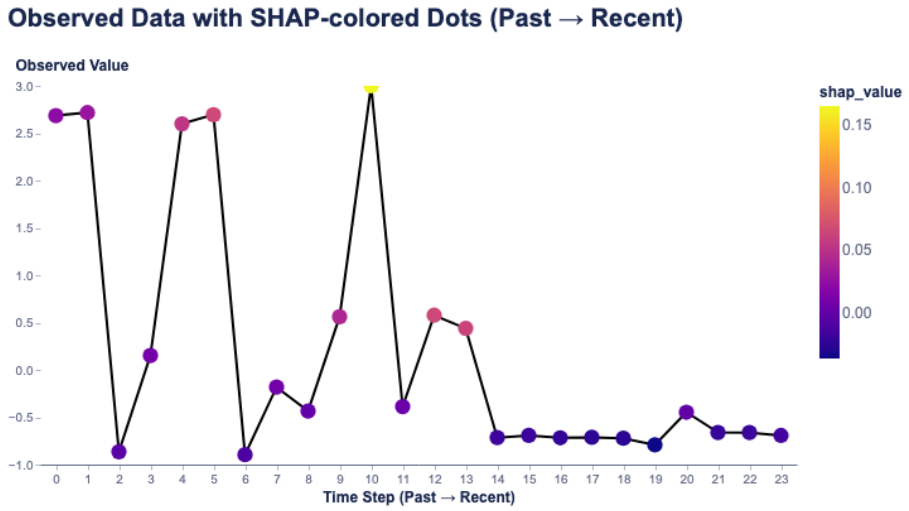

Figure 33: TimeSHAP - LSTM at test index 1000

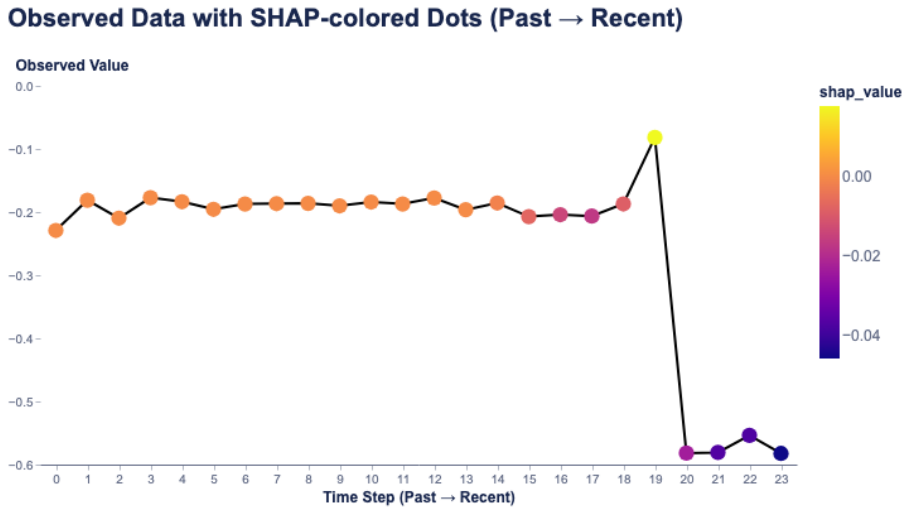

Figure 34: TimeSHAP - RNN at test index 0

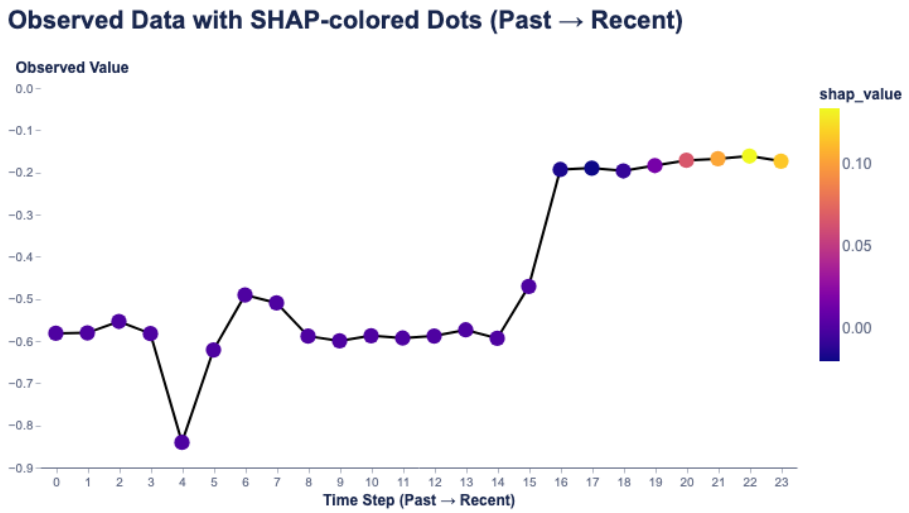

Figure 35: TimeSHAP - RNN at test index 20

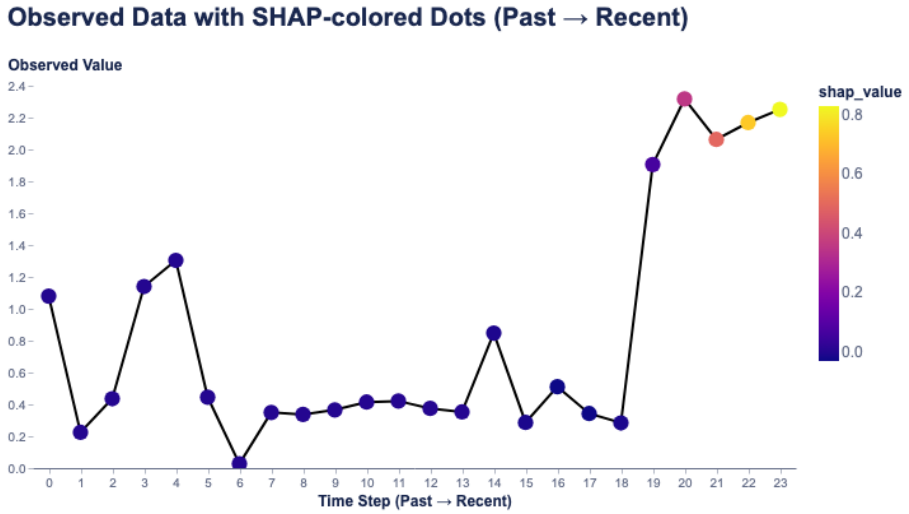

Figure 36: TimeSHAP - RNN at test index 50

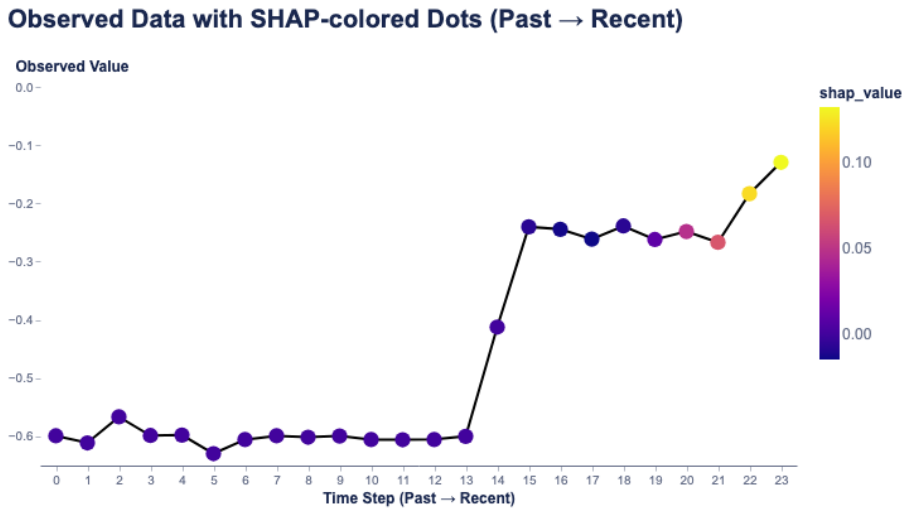

Figure 37: TimeSHAP - RNN at test index 500

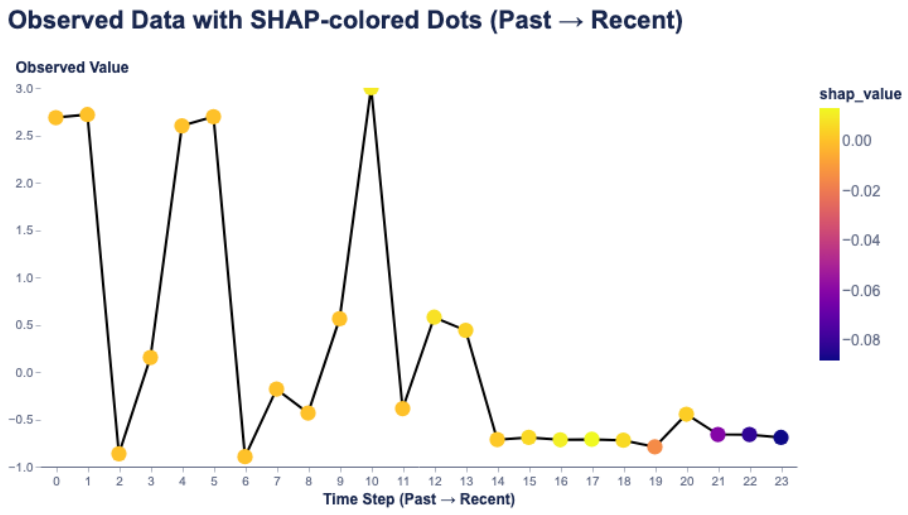

Figure 38: TimeSHAP - RNN at test index 1000