Abstract

Electricity demand forecasting is crucial for power grid planning and operations management. However, accurately predicting electricity demand remains challenging due to its dependence on diverse factors and inherent uncertainty. Electricity consumption fluctuates based on weather conditions, economic activity levels, consumer behavior and seasonal variations. The integration of renewable energy sources which have variable and intermittent outputs has further added to demand unpredictability.

Traditionally, statistical and machine learning approaches have been used for electricity demand forecasting. While these methods have improved forecasting accuracy overall, they often struggle to model extreme events and specialized contexts with limited data to train on. As a result, their predictions can become unreliable in situations that deviate from historical norms. This highlights the need for more robust and adaptive forecasting approaches that can account for exceptional conditions and complement the limitations of existing machine learning techniques, especially during periods of poor model performance.

The overarching purpose of this research is to investigate whether a human-AI collaborative approach can enhance the accuracy and robustness of electricity demand forecasting, especially during periods of poor machine learning performance. The main objectives are: (1) to evaluate if human adjustments of AI predictions can improve forecasting accuracy, particularly when machine learning models struggle. This aims to validate the potential benefit of human-AI collaboration over AI alone. (2) to gain insights into how technologically proficient users and non-technical users differ in their interaction processes and adjustment outcomes within a collaborative system. This seeks to uncover behavioral factors that influence the effectiveness of human-AI collaboration. The overall purpose is

to demonstrate the promise of human-AI collaboration for complimenting machine learning limitations through human judgment. The objectives center around testing, optimizing and proposing collaborative approaches that leverage the complementary strengths of humans and machines.

To achieve the research target, the methodology employs a mixed-methods design combining machine learning forecasting models with an interactive human-AI collaborative experiment, which has the review number 23-80 and been approved by the official ethics review committee of University of Tokyo. Historical Japanese electricity demand data alongside weather variables are selected based on analytical techniques and domain expertise. These time series are used to train a Generalized Additive Model (GAM), chosen for its flexibility in modeling nonlinear relationships. Rigorous steps are taken to tune model hyperparameters and special event selection. An interactive interface enables participants to adjust the initial statistical forecasts by leveraging their qualitative insights and understanding of contextual information. Both technical participants with AI/data science backgrounds and non-technical users participate to enable comparative analysis. Participants can make granular local adjustments or global modifications spanning the full prediction period based on perceived model limitations. Detailed behaviour logs trace every user behavior during these human-AI interactions. Lastly, the adjusted predictions are evaluated against the actual electricity demand data using quantitative error metrics like Root Mean Squared Error (RMSE). Moreover, user satisfaction ratings, adjustment rationale, and feedback surveys provide additional qualitative insights into the collaborative forecasting process. The mix of computational forecasting models and interactive human input seeks to integrate the strengths of machine learning scalability and human contextual reasoning.

The results provide quantitative evidence that there is no difference between human adjustments and statistical forecasting in general, whereas on other hand, human adjustments significantly improved model accuracy during peak demand periods (9am to 4pm), when statistical forecasting alone was inadequate due to lack of information. This finding confirms the value of incorporating human judgment into the forecasting process. Additionally, both comparative analysis and statistical testing between participant groups revealed that participants with technical backgrounds consistently produced more accurate adjustments than non-technical users. This study conducted a more in-depth analysis by examining participant behavior log data and subjective surveys after each human adjustment. The visualization analysis revealed significant differences in the operational processes of non-technical participants and participants with technical backgrounds when interacting with the forecasting system. Participants with technical backgrounds tended to be more goal-oriented, as they were able to actively search and interact with the information. On the other hand, non-technical participants performed in a more passive manner, as they were unable to build connections between different information sources. In particular, they avoided using explainable information from the GAM model. In the end, non-technical participants spent less time on the experiment and interacted with the user interface less frequently than participants with technical backgrounds. Finally, this study also provides a more detailed visualization framework that allows researchers to view the operational process and decision-making process, with probability transitions displayed. This framework may assist future research in investigating human-AI collaboration using more detailed methods.

While this thesis makes valuable contributions, certain limitations should be acknowledged. (1) Sample Size Limitations: the experiment included a relatively

small sample of 9 participants, which limits the generalizability of the results. Additionally, the participants were all recruited from one university, further reducing diversity. Expanding to larger and more diverse samples could strengthen the conclusions. (2) Task Environment Limitations: the study one forecasting task from the real-world situation. Testing on more different forecasting tasks may help to conclude the results. (3) Measurement Limitations: the subjective nature of user surveys and limited logged behavioral data poses measurement challenges. Self-reported survey are prone to biases and logged actions provide an incomplete picture. Overall, human bias is always a problem in human-AI collaboration process, especially how human receive the information and how do they process the information is completely uncertain in this research. (4) Technical Limitations: the visualization systems for tracking user behaviors require further validation to demonstrate their utility. Additionally, the computational forecasting models could likely be improved with more advanced techniques. A more friendly and intelligent user interface would strengthen the human-AI integration.

While the limitations constrain the conclusions that can drawn, the thesis still delivers valuable foundational insights. First, it provides initial empirical evidence for the benefits of collaborative forecasting frameworks that integrate human expertise with machine learning predictions. The experiments demonstrate that human adjustments, especially by those with technical knowledge, can significantly improve the accuracy of demand forecasts when statistical models alone falter. Second, the thesis proposes implementable techniques for enabling effective human-AI collaboration, including interactive interfaces, user behavior logging, coordination protocols, and survey mechanisms. These tools provide a foundation for future research into optimizing joint human-machine systems. Third, analyses of user behaviors and feedback surveys generate new insights into how

different groups interact with AI tools and key factors influencing the collaborative process. The findings highlight the need for accessible system designs that account for varying user expertise.While limited by sample size, this exploratory research lays the groundwork for further studies. Avenues for future work include scaled experiments, online deployments, control of human bias with structured collaboration methods, and more detailed and systematic methods of data collection.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Chapter 1.1 Overview

The introduction provides background information on the history and development of artificial intelligence (AI). It explains key AI concepts like machine learning and statistics, and their increasing use for forecasting tasks. In the energy industry, AI techniques have become vital for planning and sustainability efforts. However, limitations exist including data constraints, modeling extreme events, and distrust of black-box algorithms. -AI collaboration is proposed as a way to address these limitations and augment AI forecasting by incorporating human judgment. This research will examine the application of AI for electricity demand forecasting in Japan, with a focus on developing and evaluating collaborative forecasting approaches that combine AI predictions with human inputs. The goal is to improve forecasting accuracy by leveraging the complementary strengths of both machine and human intelligence.

Chapter 1.2 Background

Artificial Intelligence (AI) is a broad field of study that encompasses various subfields and techniques. It is often described as the simulation of human intelligence processes by machines, especially computer systems. These processes include learning, reasoning, problem-solving[i], perception, and language understanding (Russell & Norvig, 2020).

The origins of AI research can be traced back to the 1940s and 1950s, when pioneering researchers like Alan Turing, John McCarthy, Marvin Minsky and Claude Shannon laid the foundations for intelligent machines (Nilsson, 2010). Some key developments in those early days included Turing's formulation of the Turing Test to evaluate machine intelligence, McCarthy's coining of the term "artificial intelligence", and Minsky and Shannon's work on neural networks (Russell & Norvig, 2020). In the 1960s and 1970s, AI researchers focused on solving general reasoning and knowledge representation problems. Approaches like expert systems, natural language processing, robotics and machine learning were explored. After 1980s, public saw the rise of commercial AI applications, but funding dried up later as limitations of the technology became apparent, leading to the "AI winter". In the 1990s and 2000s, AI slowly regained prominence through more specialized applications using machine learning and probabilistic methods. The defeats of chess champion Garry Kasparov by IBM's Deep Blue and quiz show contestants by Watson indicated major advances (Copeland, 2022).

The 2010s and beyond marked the resurgence of AI through explosive growth in big data, advanced algorithms like deep learning, and increased computing power via GPUs and cloud computing. AI began finding extensive use in specialized tasks like computer vision, speech recognition and machine translation (Jordan & Mitchell, 2015).

Now the definition of Artificial Intelligence is (Samoili et al., 2020):

"***Artificial Intelligence refers to systems that display intelligent behaviour by analysing their environment and taking action — with some degree of autonomy — to achieve specific goals.***".

The primary goal of AI is to create systems that can function intelligently and independently. AI aims to create machines that can mimic human intelligence and perform tasks that would normally require human intellect. These tasks include understanding natural language, recognizing patterns, solving complex problems, and support in decisions-making. Natural language understanding refers to the ability to parse, process and generate human language. This allows AI systems to analyze text or speech data to perform tasks like translation, summarization and question answering (Jordan & Mitchell, 2015). Pattern recognition involves identifying patterns in data that can be used for classification or prediction. AI uses pattern recognition for facial recognition, product recommendation, financial fraud detection etc (Russell & Norvig, 2020).

Machine learning plays a crucial role in the field of AI. It is the study of computer algorithms that improve automatically through experience and by the use of data (Alpaydin, 2020). Machine learning techniques like neural networks, reinforcement learning and deep learning have been applied successfully in areas like computer vision, Natural Language Processing (NLP), robotics, and bioinformatics (James et al., 2013).

Statistics play a significant role in AI. It provides the foundation for machine learning and AI, allowing for the understanding and quantification of uncertainty (James et al., 2013). Statistical methods like Bayesian inference, regression, principal component analysis are often used in AI for tasks such as parameter estimation, hypothesis testing, dimensionality reduction, and correlation analysis (James et al., 2013).

AI has been increasingly applied across multiple sectors beyond healthcare. In finance, AI techniques are used for fraud detection, credit scoring, algorithmic trading, and personalized banking (Ngai et al., 2011). For supply chain management, AI enables demand forecasting, inventory optimization, and predictive maintenance (Ivanov, Dolgui, & Sokolov, 2019). Other applications include autonomous vehicles, cybersecurity, education, manufacturing, and more (Kaplan & Haenlein, 2019). In the medical field, AI has been used for medical imaging analysis, robotic surgery, virtual nursing assistants, and drug discovery (Jiang et al., 2017). AI also played a key role during the COVID-19 pandemic in areas like viral genome sequencing, CT scan diagnosis, vaccine development, and patient prioritization (Vaishya et al., 2020).

In the energy sector, AI techniques have been used to improve efficiency, planning and sustainability. AI has been applied to various aspects of the energy industry including demand forecasting, asset maintenance, power grid optimization and integrating renewable energy through smart grids (Mhlanga, 2023; Entezari et al., 2023).

AI offers numerous benefits and advantages over traditional methods. It provides opportunities for competitive advantages and has a significant impact on society and firms (Von Krogh, 2018). AI can improve processes and often benefits from cognitive work redesign efforts to augment human capabilities (Daugherty &

Wilson, 2018). However, there are also concerns around biases, privacy, security, ethics and social impacts that should be considered with the rapid rise of AI.

Forecasting is one of the main applied fields in artificial intelligence field, forecasting refers to the process of making predictions about future events based on historical data and patterns (Armstrong, 2001). With recent advances in artificial intelligence (AI), forecasting techniques leveraging machine learning have become increasingly popular and effective, especially for time series forecasting (Makridakis, Spiliotis & Assimakopoulos, 2020). Time series forecasting utilizes observations over time to uncover temporal relationships and make probabilistic forecasts about the future (Hyndman & Athanasopoulos, 2018). It has wide applications across domains like finance, energy, and transportation.

Several algorithms have been developed for time series forecasting. Traditional statistical methods include autoregressive models like ARIMA that rely on lagged relationships (Liu, 2022), and regression techniques like generalized linear models (GLMs) and generalized additive models (GAMs) that incorporate explanatory variables (Brockwell & Davis, 2016). More advanced machine learning techniques include random forests (RF), which aggregate decisions from multiple decision trees, and long short-term memory (LSTM) neural networks that can uncover longer-term dependencies (Punia, 2020). Hybrid methods like Prophet combine traditional statistics with machine learning for greater flexibility and accuracy (Taylor & Letham, 2018).

AI has become invaluable in the electricity industry for forecasting demand, pricing, and supply needs (Qdr, 2006). In the case of Japan, for example, Japan has undertaken extensive electricity reform and aims to derive 22-24% of its energy from renewable sources by 2030 (METI, 2018). However, fluctuating renewable energy makes demand forecasting challenging (Sugihara et al., 2012). AI

techniques like LSTM networks can uncover complex data patterns to improve short-term load forecasts (Ageng, Huang & Cheng, 2021). But limitations remain regarding sparse data, unprecedented events, and social factors like distrust in AI (Bengio, Courville &Vincent, 2013; Amodei et al., 2016).

By solving the problem generally exists in artificial forecasting algorithms, Human-AI collaboration (HACI) refers to combining the predictions of AI systems with human judgment and domain knowledge to arrive at augmented forecasts (Patrick, 2023). This can leverage the benefits of both AI's ability to uncover complex data patterns with human context and reasoning skills. There is increasing need for such collaboration in complex domains like energy forecasting where data may be limited and social factors come into play (Fast & Horvitz, 2017). This research will examine how incorporating human input into an AI-generated electricity demand forecast can enhance accuracy, especially in cases where machine learning alone may falter.

Chapter 1.2 The organization of the paper

This paper is organized into eight chapters to give a comprehensive analysis of the application of artificial intelligence (AI) in electricity forecasting, the challenges associated, and the innovative solutions that integrated with human factor in forecasting process.

Chapter 1 introduces the topic and provides an overview and background to the study, after which the organization of the paper is explained. In Chapter 2, a literature review is conducted on various topics, including the definition and application of AI, particularly in forecasting. The chapter further delves into traditional and machine learning methods of forecasting, their comparison, and

their specific use in electricity forecasting. The mechanism and limitations of current forecasting algorithms are discussed, ending with a section on human and AI collaboration. Chapter 3 presents the problem statement, including an overview and an exploration of factors influencing electricity forecasting. We then discuss the current challenges in this field, with particular focus on general challenges and those unique to Japan. The role of AI and the need for human-AI collaboration in Japanese electricity demand forecasting are also analyzed in this chapter.In Chapter 4, the research methodology is outlined, starting with an overview, followed by a detailed exposition of the process that includes data collection, processing, analysis, and discussion on the current forecasting method in Japan. This chapter also introduces the General Additive Model (GAM) for forecasting and explains the model selection and training, ending with a review of evaluation methods. Chapter 5 discusses human collaboration in forecasting, elaborating on the integration of the human factor, judgmental forecasting, and the risks, advantages, and challenges associated with it. It also outlines the types of judgmental forecasting methods and how to design an effective Human-AI collaboration. This chapter introduces the experiment design as well. Chapter 6 presents the experiment design, outlining its objectives, participant group, training, procedures, data collection, analysis, and the expected results. In Chapter 7, the results and discussion are presented, including an overview, a comprehensive analysis of the collected data, a comparison of human adjusted results and pure machine predictions. The chapter ends with an in-depth analysis of the main results and the data concerning the Human-AI collaboration. Finally, Chapter 8 concludes the paper by summarizing the main findings and drawing overall conclusions from the research.

Chapter 2 - Literature Review

Chapter 2.1 Overview

This literature review examines research on artificial intelligence (AI) forecasting methods and human-AI collaboration in forecasting.

The review begins by defining key terminology around AI and machine learning. It then provides an overview of traditional statistical forecasting methods as well as more advanced machine learning techniques that have been applied to electricity demand forecasting. The comparative advantages and limitations of different forecasting methods are discussed. Relevant research on forecasting methods in the Japanese context is also summarized.

The mechanics behind the forecasting process are outlined, including important steps like data collection, preprocessing, feature selection, model training/testing, and post-modeling procedures. Key challenges and limitations of current AI forecasting algorithms are also highlighted, such as data constraints, interpretability issues, evaluation difficulties, and modeling of extreme events.

Finally, the emerging field of human-AI collaboration is introduced. The potential benefits of combining human and AI capabilities for improved forecasting performance is discussed. However, open challenges around trust, transparency, coordination, and communication in human-AI teams are also acknowledged. Effective training and interface design are noted as important factors for successful collaboration.

In summary, this review covers the landscape of AI techniques for forecasting while also identifying limitations and gaps that provide opportunities for further research. It lays the groundwork for proposing and evaluating collaborative forecasting approaches that harness the complementary strengths of human experts and AI systems.

Chapter 2.2 What is artificial intelligence (AI) & the definition of AI

First of all, artificial intelligence (AI) refers to intelligent systems that can analyze environments, take autonomous actions, and achieve goals (European Parliamentary Research Service, 2020). However, industry and academia often categorize AI differently, leading to confusion between terms like AI, machine learning (ML), and deep learning (DL) (Tiwari, Tiwari & Tiwari, 2018). At the beginning of this thesis, a clear definition of artificial intelligence (AI) is important.

Artificial Intelligence (AI) has come a long way since its inception, with its initial focus being on logical reasoning. The field has expanded its capabilities with the advancement of computing power, enabling machines to perform tasks that were once thought to be exclusive to humans (Russell & Norvig, 2016). The early days of AI were characterized by a focus on logical reasoning. This was a time when the capabilities of AI were largely theoretical, and the practical applications were limited. However, the advancement of computing power allowed for an expansion of these capabilities, leading to the development of more complex AI systems (Russell & Norvig, 2016). The 1960s and 70s marked a significant period in the development of AI, with the creation of expert systems. These systems were designed to mimic the decision-making abilities of human experts, providing solutions to complex problems in various fields such as medicine and engineering (Buchanan, 2005). The 1980s saw the rise of commercial applications of AI. However, this period was followed by what is often referred to as the "AI winter", a time of reduced funding and interest in AI research due to the limitations of the technology and inflated expectations (McCorduck, 2004). The 1990s marked the beginning of a revival in AI research and development, with a focus on specialized domains such as computer vision and natural language processing. The increase in

data and computing power has led to significant advancements in these areas, enabling AI to perform tasks with a level of sophistication that was previously unimaginable (Goodfellow, Bengio, & Courville, 2016). Despite the significant advancements in AI, there are still notable limitations. One of the main challenges is the lack of common sense in AI systems. Additionally, the performance of AI systems is heavily dependent on the availability and quality of data. These limitations highlight the need for ongoing research and development in the field of AI (Vaishya, Javaid, Khan, & Haleem, 2020).

Machine learning (ML) and deep learning (DL) are widely used subsets within the broader field of AI. ML focuses on algorithms that can learn from data, while DL uses neural networks and massive data for complex tasks like computer vision and NLP (Alpaydin, 2020). Besides techniques, AI also comprises goals like classification and forecasting, and specialized tasks like robotics, analytics, and more.

The multifaceted nature of AI has led to ambiguity, multiplicity, and subjectivity in its definition (Wang, 2019). Ambiguity arises from the diverse capabilities of AI. Multiplicity refers to the broad application of AI across industries and tasks. Subjectivity deals with differing perspectives on the degree of intelligence exhibited by AI systems. These challenges make it difficult to establish a unified definition. In his work, his research indicates that there is no confusion in understand of A (artificial), but the understand of I (intelligence) varies largely according to the historical stage and context.

This study adopts the view that intelligence represents the ability to adapt to environments given insufficient knowledge and resources (Wang, 1995).

"Intelligence is the capacity of an information processing system to adapt to its environment while operating with insufficient knowledge and resources."

Specifically, we define AI as using statistical models for time series forecasting, where accuracy in predicting outcomes is the key metric of evaluation. The application domain is energy industry forecasting, which exemplifies AI constraints on data, computing power, and expertise. As McCarthy (2007) stated:

"AI is concerned with methods of achieving goals in situations where information is complex."

This study has the same idea with this concept. This study believes that AI strictly solves problems suitable for different environments through information processing capabilities. Here, the environment is a broad term representing the goals of solving problems, that is, different industries, scenarios, etc. Finally, it is necessary to clarify how to understand "insufficient knowledge and resources." This study believes this is the best supplement to understanding AI. First, in the eyes of the concept proposer, the normal working state of AI is not an unlimited knowledge storage space, unlimited computing resources or unlimited computing time, but limited or even scarce. In the practical application of AI, we believe this is a very important assumption because data acquisition and computing resources are limited. Although AI development is gradually breaking through the limits of computing power and parameters with the increase in computing power and the proposal of large language models, it does not mean that the dilemma of limited resources has been generally solved.

Chapter 2.3 Artificial intelligence in forecasting

For the definition of AI in forecasting from the research paper "Judgmental Adjustments in Demand Forecasting: A Review of Progress over the Last 25 Years" by Lawrence et al. (2006):

> ***"In forecasting, artificial intelligence (AI) refers to the use of computer-based techniques to model and predict future events. AI techniques can be used to improve the accuracy of forecasts by identifying patterns in historical data that would otherwise be missed. AI techniques can also be used to incorporate human judgment into forecasting models, which can help to improve the accuracy of forecasts in situations where the data is not representative of the real-world problem."***

Artificial intelligence (AI) techniques can improve forecast accuracy by identifying patterns in data that may be missed by traditional statistical methods. AI can also integrate human judgment to account for real-world complexities not captured in historical data (Lawrence et al., 2006). This section reviews common forecasting techniques for electricity demand.

Author(s)	Methods	Findings
Shah et al. (2020)	Parametric and Nonparametric Approaches	Effective for one-day-ahead electricity price forecasting in the Italian market
Yang et al. (2013)	New Strategy for Short-term Load Forecasting	Improved forecasting accuracy
Alharbi and Csala (2022)	SARIMAX Forecasting Model-based Time Series Approach	Effectively captured seasonal patterns in electricity demand
Setiyorini and Frieyadie (2020)	Comparison of Linear Regressions and Neural Networks	Neural networks outperformed linear regressions in forecasting electricity consumption

Fan and Hyndman (2012)	Semi-parametric Additive Model	Addressed limitations of linear regression methods in short-term load forecasting
Marino et al. (2016)	Deep Neural Networks (DNNs)	Effective for building energy load forecasting
Yuce et al. (2017)	Smart Forecasting Approach using Machine Learning	Effective for district energy management
Khafaf et al. (2019)	Long Short-Term Memory (LSTM) Networks	Effective in capturing complex patterns in energy demand data
Lee and Cho (2022)	Traditional, Machine Learning, or Hybrid Model	Machine learning models outperformed traditional models in national-scale electricity peak load forecasting
Makridakis et al. (2018)	Comparison of Statistical and Machine Learning Forecasting Methods	Raised concerns about over-reliance on machine learning forecasting methods
Li and Lu (2023)	GWO-LSTM Network	Outperformed traditional methods in short-term power forecasting
Jiang et al. (2019)	Modeling of Electricity Demand Forecast for Power System	Machine learning methods provided more accurate forecasts
Otsuka (2016)	Determinants of Residential Electricity Demand	Income, price, and temperature were significant factors
Otsuka (2019)	Case Study on the Impact of the 2011 Great East Japan Earthquake	Significant impact on electricity consumption behavior
Elamin and Fukushige (2018)	SARIMAX Model with Interactions	High accuracy in hourly demand forecasting
Zhang et al. (2012)	Scenario Analysis on Future Electricity Supply and Demand	Considered factors such as economic growth, energy efficiency, and energy policies

Table 2.1 The overview of forecasting methods

Table 2.1 summarizes key studies covered in the literature review that have applied different forecasting methods to electricity demand prediction tasks. The table is organized by listing the author(s) of each study, the specific forecasting methods or models they utilized, and the major findings or results highlighted from their research.The methods encompass both traditional statistical approaches like SARIMAX models as well as more advanced machine learning techniques like deep neural networks and LSTM networks. The key findings demonstrate the effectiveness of different techniques for tasks like short-term load forecasting,

capturing seasonal patterns, predicting extreme events, and scenario analysis of future supply and demand. Some studies also provide comparative insights, such as machine learning models outperforming traditional methods in peak load forecasting, or neural networks showing superior accuracy over linear regression models in electricity consumption prediction. The table highlights the diversity of forecasting techniques applied in recent literature while also condensing the key takeaways and performance outcomes associated with each method.

In the next section, this research will provide a more in-depth discussion of forecasting methods.

Chapter 2.3.1 Traditional forecasting methods

Traditional methods for electricity demand forecasting primarily relied on statistical techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA) (Taylor & McSharry, 2007), and Seasonal ARIMA with exogenous variables (SARIMAX). These models have been effective in capturing linear relationships in the data. However, they have limitations in capturing nonlinear relationships, which are often present in electricity demand data due to factors such as weather conditions and time of day (Taylor, 2003). To address these limitations, advanced statistical models like Generalized Additive Models (GAM) have been used for load forecasting (Wood, 2006).

In traditional methods, Shah et al. (2020) used parametric and nonparametric approaches to forecast one-day-ahead electricity prices for the Italian electricity market. Yang et al. (2013) proposed a new strategy for short-term load forecasting, which significantly improved the forecasting accuracy. Alharbi and Csala (2022) developed a SARIMAX forecasting model-based time series approach, which effectively captured the seasonal patterns in electricity demand. However, these

traditional methods have limitations in capturing nonlinear relationships, which are often present in electricity demand data. For instance, Setiyorini and Frieyadie (2020) found that neural networks outperformed linear regressions in forecasting electricity consumption. Fan and Hyndman (2012) proposed a semi-parametric additive model for short-term load forecasting, which addressed the limitations of linear regression methods.

Chapter 2.3.2 Machine learning methods

With the advent of machine learning, new methods for electricity demand forecasting have been developed. These include deep learning architectures such as Deep Neural Networks (DNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) (Kong et al., 2017). These models have shown superior performance in complex electricity forecasting tasks due to their ability to account for multivariate, nonlinear relationships (Goodfellow et al., 2016)

These methods include deep learning architectures such as Deep Neural Networks (DNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs). Marino et al. (2016) used DNNs for building energy load forecasting and achieved promising results. Yuce et al. (2017) developed a smart forecasting approach to district energy management using machine learning methods. Khafaf et al. (2019) applied LSTM in energy demand forecasting and found it to be effective in capturing complex patterns in the data. Lee and Cho (2022) conducted a national-scale electricity peak load forecasting study and found that machine learning models outperformed traditional models.

.

Chapter 2.3.3 Comparison of different methods

Machine learning models have several advantages over traditional statistical models. They are capable of dynamically modeling interactions between variables, which can lead to more accurate forecasts. In a comparison study, machine learning models were found to outperform ARIMA models for peak load forecasting (Cancelo et al., 2008). Makridakis et al. (2018) raised concerns about the over-reliance on machine learning forecasting methods and emphasized the importance of understanding their limitations. Li and Lu (2023) developed a short-term power forecasting model based on the GWO-LSTM network, which outperformed traditional methods. Jiang et al. (2019) modeled electricity demand forecast for power systems and found that machine learning methods provided more accurate forecasts

In the context of Japan, Otsuka (2016) studied the determinants of residential electricity demand and found that income, price, and temperature were significant factors. Otsuka (2019) conducted a case study on the impact of the 2011 Great East Japan Earthquake on electricity consumption behavior. Elamin and Fukushige (2018) used a SARIMAX model with interactions for hourly demand forecasting and achieved high accuracy. Zhang et al. (2012) conducted a scenario analysis on future electricity supply and demand in Japan, considering factors such as economic growth, energy efficiency, and energy policies.

Chapter 2.3.4 Methods of forecasting electricity

Several methods have been developed for forecasting electricity demand. These include partial adjustment models for estimating electricity demand functions, price elasticity analysis for understanding demand fluctuations, and the SARIMAX

model with interactions for hourly demand forecasting (Taylor, 2010). Scenario analysis has also been used to forecast future electricity supply and demand, taking into account factors such as economic growth, energy efficiency, and energy policies (Hong et al., 2016). For instance, studies have predicted an expected increase in electricity demand in Japan and emphasized the importance of energy efficiency and renewable energy sources (Akimoto et al., 2010)

Chapter 2.4 Mechanism of forecasting

This section provides an overview of the core components and workflow of an AI forecasting system. Key steps include data collection and preprocessing, feature selection, model training/testing, model selection and evaluation, and finally model deployment and updating. Understanding this end-to-end forecasting pipeline is important for developing accurate AI prediction models. The subsequent sections explain each step in detail, highlighting important concepts and techniques used in practice. Overall, this overview summarizes the systematic methodology needed to transform raw data into trained models that can generate reliable forecasts.

❖ Data Collection

Data collection is the initial step in AI forecasting. It involves gathering relevant data that will be used to train the forecasting models. The quality and quantity of the data collected significantly influence the accuracy of the forecast (Kelleher, Mac Namee, & D'Arcy, 2015).

❖ Data Preprocessing

Data preprocessing is a crucial step in AI forecasting. It involves cleaning and transforming raw data into a format that can be easily understood and used by

machine learning algorithms. This step is necessary because real-world data is often incomplete, inconsistent, and noisy (García, Luengo, & Herrera, 2015).

- ✓ Handling Missing Values: Missing values in the dataset can lead to inaccurate forecasts. Various techniques such as imputation can be used to handle missing values (Luengo, García, & Herrera, 2012).
- ✓ Outlier Detection and Removal: Outliers can skew the model's understanding of the data, leading to inaccurate forecasts. Techniques such as the Z-score method can be used to detect and remove outliers (Zhang, 2016).
- ✓ Normalization: Normalization is the process of scaling numeric data from different input variables down to a similar scale. This process can make training less sensitive to the scale of features, allowing the model to converge faster (Patro & Sahu, 2015).

❖ Feature Selection

Feature selection is the process of selecting the most relevant features for use in model construction. It reduces the dimensionality of the data and enables the learning algorithm to operate faster and more effectively (Chandrashekar & Sahin, 2014).

- ✓ Methods in Machine Learning: Common methods include filter methods, wrapper methods, and embedded methods. Each has its advantages and disadvantages, and their use depends on the specific requirements of the forecasting task (Saeys, Inza, & Larrañaga, 2007).
- ✓ Methods in Statistics: Techniques such as stepwise selection and backward elimination are commonly used. These methods have their strengths and

weaknesses and are chosen based on the data and the specific needs of the analysis (Heinze, Wallisch, & Dunkler, 2018).

❖ Splitting Data

Data splitting is the process of dividing the dataset into two or more subsets. The purpose of this is to assess the performance of the model on unseen data. The subsets include: 1)Training Set: This is the subset of the data on which the model is trained. 2) Testing Set: This is the subset of the data on which the model is tested to evaluate its performance (Kohavi, 1995).

❖ Post-modeling procedure

Model selection involves choosing the best model from a set of candidate models based on their performance. In AI forecasting, several models can be used, including ARIMA, Exponential Smoothing, Prophet, RNNs, CNNs, and Ensembles. The selection of the model depends on the nature of the task and the data (Christoffersen & Diebold, 1996). Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent dataset. It is primarily used in settings where the goal is to predict future outcomes on the basis of other related information. It's commonly used in time-series forecasting with methods like forward chaining where you model on past data then look at forward-facing data (Bergmeir, Hyndman, & Koo, 2018). Model selection involves choosing the best model from a set of candidate models based on their performance. In AI forecasting, several models can be used, including ARIMA, Exponential Smoothing, Prophet, RNNs, CNNs, and Ensembles. The selection of the model depends on the nature of the task and the data (Christoffersen & Diebold,

1996). Model evaluation is the process of determining how well a model performs on unseen data. It involves calculating various metrics that quantify the difference between the predicted and actual values. Common metrics include RMSE, MAPE, sMAPE, and MAE. Other approaches include out-of-sample (OOS) approaches, prequential approaches, and cross-validation approaches (Shcherbakov et al., 2013).

After a model has been trained and evaluated, it is deployed to make forecasts. This involves using the model to predict future values based on new input data. The model may also need to be updated or retrained over time as new data becomes available (Raza & Khosravi, 2015).

Chapter 2.5 The limitation of current forecasting algorithm

These machine generated algorithms, often based on machine learning techniques, have the potential to predict future events or trends with high accuracy. However, they are not without limitations. Understanding these limitations is crucial in this research to understand the potential problems and solutions to them. (Dineva & Atanasova, 2020).

The first limitation is data limitations. Many advanced forecasting algorithms like deep neural networks and random forests rely on large, clean datasets to accurately model complex patterns (Xu et al., 2023). However, acquiring sufficient high-quality data can be challenging in practice. Issues like missing data, noise, biases, and errors in datasets can significantly degrade model performance (Petropoulos et al., 2022). Algorithms trained on limited or flawed data often fail to generalize well to new contexts (Vabalas et al., 2019). Data pre-processing and augmentation techniques can help, but do not fully mitigate the need for large datasets representative of the true environment.

The second limitation comes from the interpretability of the model. There are inherent tradeoffs between accuracy and interpretability for many complex yet powerful forecasting models like deep neural networks (Moss et al., 2022). The internal workings of such black-box models are not easily understood, making it difficult to diagnose failures or identify biases. Additionally, many models struggle to effectively predict rare or extreme events not well-represented in training data (Ding et al., 2019). Capturing complex nonlinear and dynamic relationships poses challenges for simpler forecasting algorithms with linear assumptions (Petropoulos et al., 2022). With limited data, overfitting can easily occur, reducing generalizability (Vabalas et al., 2019). Algorithm aversion is another phenomenon where people fail to use algorithms after learning that they are imperfect. Dietvorst et al. (2018) discuss how people are more likely to use an imperfect algorithm if they can modify its forecasts, suggesting a desire for control over the forecasting process. Similarly, Burton et al. (2019) and Islam et al. (2022) discuss how algorithm aversion can impact decision-making, with people often preferring less accurate human judgment over more accurate algorithmic predictions.

The third limitation is about evaluation difficulties. Common forecast error metrics like root mean squared error (RMSE) have limitations and may not fully reflect real-world costs of inaccurate predictions (Petropoulos et al., 2022). Rigorous validation procedures like nested cross-validation are computationally expensive, often prohibiting their use in practice due to time constraints (Vabalas et al., 2019). It is also difficult to simulate actual forecasting environments and workflows, meaning models may degrade when deployed in operational systems.

Lastly, which is most important challenge that this research thinks in reality. Modeling extreme events in time series prediction is a significant challenge. Ding et al. (2019) discuss how deep learning methods often overlook extreme events,

leading to poor performance in real-world time series. They propose a new form of loss function and a memory network module to improve the modeling of extreme events.

In conclusion, while forecasting algorithms have significant potential, they also have numerous limitations. These include the risk of overfitting, the need for large amounts of data, difficulties with interpretability and trust, and challenges with modeling extreme events. Further research is needed to address these limitations and improve the effectiveness of forecasting algorithms.

Chapter 2.6 Human and AI collaboration

Human-AI collaboration (HAIC) is a rapidly evolving field that aims to create synergistic teaming between human decision-makers and AI systems, leveraging the strengths of both to optimize performance and fairness in decision-making processes (Leitão et al., 2022). This collaboration is increasingly becoming a part of organizational decision-making, with AI technologies being integrated into strategic decision-making processes (Trunk et al., 2020). It is characterized by the interaction between humans and AI in a shared task, where AI systems are designed to learn from, adapt to, and work with human partners (Jarrahi, 2018).

In the context of design, human-AI collaboration involves the use of AI technologies to support and enhance human creativity and problem-solving abilities (Dellermann et al., 2019). This collaboration has also been demonstrated in the medical field, where AI systems have been used to assist human decision-making, leading to improved diagnostic accuracy (Reverberi et al., 2022). Moreover, in healthcare, human-AI collaboration is seen as a promising approach to improve patient care and health outcomes (Lai et al., 2021).

However, the success of human-AI collaboration depends on several factors, including the design of the AI system, the management of human-AI interactions, and the transparency of the AI system (Abedin et al., 2022; Vössing et al., 2022). Trust is also a critical factor in human-AI collaboration, as it influences the willingness of humans to rely on AI systems (Bao et al., 2021). Training is essential to foster effective human-AI collaboration, as it helps humans understand the capabilities and limitations of AI and how to best interact with it (Figoli et al., 2022).

Furthermore, the concept of trustworthy human-AI collaboration has been introduced, which emphasizes the need for AI systems to be reliable, ethical, and transparent to foster trust and effective collaboration with humans (Razmerita & Brun, 2022). Despite the potential of human-AI collaboration, there are challenges and limitations that need to be addressed, such as the requirements of learning to defer (L2D) framework and the issues of deploying HAIC systems in real-world settings (Leitão et al., 2022).

Advantages/Disadvantages of Machine Learning: Machine learning offers several advantages, including the ability to process large volumes of data, identify complex patterns, and make predictions with high accuracy, which is beyond human capabilities (Dineva & Atanasova, 2020). However, it also has its disadvantages. Machine learning algorithms often struggle with small data sets and the prediction of extreme events, which can lead to inaccurate predictions (Onyema et al., 2022; Ding et al., 2019). Furthermore, the "black box" nature of many machine learning algorithms can lead to trust issues, as users may not understand how the algorithm arrived at a particular decision (Nakashima et al., 2022).

Advantages/Disadvantages of Humans: Humans have the advantage of understanding context, making sense of ambiguous situations, and applying ethical considerations, which are areas where AI currently falls short (Trunk et al., 2020). However, humans are also subject to cognitive biases and have limitations in processing large amounts of data, which can lead to errors in decision-making.

Advantages/Disadvantages of Collaboration: The collaboration between humans and AI can leverage the strengths of both, leading to improved decision-making (Jarrahi, 2018). It allows for the combination of human intuition and AI's computational power, potentially leading to better outcomes than either could achieve alone (Dellermann et al., 2019). However, collaboration also brings challenges, such as the need for effective communication and understanding between humans and AI, and the risk of over-reliance on AI.

Despite the promising outlook, several challenges remain in implementing effective human-AI collaborative systems. A primary issue is establishing trust between humans and AI agents. Bao et al. (2021) highlight that lack of transparency into AI reasoning leads to distrust. Vössing et al. (2022) empirically showed transparency positively influenced trust and task performance in human-AI teams. Other challenges include coordination difficulties, communication barriers, role conflicts, and ambiguity in responsibility assignment between humans and AI (Abedin et al., 2022). Humans may misunderstand AI behaviors leading to improper reliance, underutilization or overdependence (Lai et al., 2021). Biases in data and algorithms also raise concerns about unfairness (Suresh & Guttag, 2021).

As AI capabilities grow, human-AI collaboration systems are expected to become ubiquitous across sectors. Key emerging trends in this direction include natural language interfaces enabling seamless coordination between humans and

AI, and networked AI agents collaborating amongst themselves and with multiple humans (Stone et al., 2016). Advances in context awareness, ethics and emotion recognition will allow AI systems to function more naturally in hybrid teams. However, challenges around transparency, accountability and fairness will continue to require research, it is important to consider those challenge in this research.

Chapter 3 - Problem Statement

Chapter 3.1 Overview

The Japanese electricity industry has undergone significant reform since the 1990s to transition from a system of regional vertically integrated utility monopolies to a more liberalized and competitive structure. The reform process began in 1995 with the introduction of independent power producers in the generation sector (JEPIC, 2023). This was followed by a staged opening up of the retail market, starting with high-demand industrial users in 2000 and expanded to full retail competition by 2016 (JEPIC, 2023).

Additional major reform steps included the establishment in 2015 of the Organization for Cross-regional Coordination of Transmission Operators (OCCTO) to coordinate grid operations nationwide (JEPIC, 2023), and the legal separation in 2020 of the transmission/distribution businesses of the former monopoly utilities from their generation and retail operations (JEPIC, 2023). The goal of these market reforms has been to improve efficiency and enable consumer choice while also ensuring energy security, integrating renewables, and providing universal service (Ohashi, 2010). However, the former regional utilities still hold a dominant market position, regulated retail rates remain in place for households, and challenges around issues like capacity adequacy persist (Shinkawa, 2022).

This chapter will provide an overview of the current structure of Japan's electricity industry and a discussion of key developments in the market reform process. Topics covered include the licensing system for operators in the generation, transmission/distribution and retail sectors, the roles of regulatory bodies like the Electricity and Gas Market Surveillance Commission, an outline of the wholesale and retail electricity markets, and an update on efforts to further liberalize the market. These market dynamics and policy shifts have significant impacts on electricity demand patterns and uncertainty.

The chapter further examines diverse forecasting influences including weather, economics, demographics, social factors, and renewable generation intermittency. Particular complexities in forecasting electricity demand in Japan are highlighted, like vulnerability to natural disasters and global energy supply shocks. Given these multifaceted challenges, the chapter argues for necessary human-AI collaboration in electricity demand forecasting. While AI offers scalable data processing and pattern identification capabilities, human expertise provides critical context on societal factors and oversight on model outputs. However, effective human-AI coordination remains a challenge, along with model interpretability limitations and bounds on expertise. Overall, blended human-AI capabilities are presented as essential for producing reliable forecasts tailored to Japan's unique landscape. The discussion sets the stage for evaluating collaborative approaches to meet the country's needs.

Chapter 3.2 Factors that can influence the electricity forecasting

This section discusses various factors that influence electric load forecasting, including meteorological factors, temporal factors, calendar factors, economic factors, customer factors, random factors, and other factors. Understanding the impact of these factors and their degree of influence is essential for enhancing the accuracy and reliability of load forecasting models, which in turn supports efficient power system planning and distribution (Fahad & Arbab, 2014; Khatoon et al., 2014).

Factor	Description	Influence on Electricity Demand
Temperature	Air temperature impacts heating and cooling requirements	Higher temperatures increase demand for cooling/AC. Lower temperatures increase heating needs.
Humidity	Higher humidity intensifies the effects of high/low temperatures	Can increase electricity demand for cooling in summer
Wind Speed	Cools the air temperature, speeding up heat loss from skin	Cools apparent temperature, reducing cooling electricity use in summer
Precipitation	Rain and snow affect air temperature	Lower demand in summer, higher demand in winter
Cloud Cover	Blocks sunlight, reducing cooling effect in daytime	Lower daytime demand in summer, higher demand in winter
Time of Day	Electricity use varies based on daily human cycles	Peaks in morning, evening, dips overnight
Day of Week	Weekday vs weekend activity patterns differ	Weekdays higher demand from offices/industry
Season	Weather impacts from seasonal cycles	Higher cooling demand in summer, higher heating in winter
Holidays	Changes activity patterns	Lower industrial/commercial electricity use
Economy	GDP, income levels, electricity pricing	Higher economic development increases overall demand
Random Events	Sudden load spikes from factories, weddings etc	Hard to predict, can cause volatility in demand
Location	Urban vs rural, access to AC	Urban areas higher demand, increased AC access raises demand
Customer Type	Residential, commercial, industrial	Distinct demand patterns for each customer class

Table 3.1 Factors that have impact on electricity demand

Meteorological factors play a significant role in electricity forecasting, particularly for renewable energy systems such as wind and solar power. Wind turbines rely on wind speed to generate electricity, making accurate meteorological data essential for forecasting the potential energy production from wind farms. Similarly, solar power generation depends on the availability of sunlight or solar irradiance. Accurate forecasts of solar irradiance enable better predictions of solar power generation. Temperature also affects electricity demand patterns, with hot weather increasing demand for air conditioning and cold weather raising heating requirements (Fahad & Arbab, 2014).

Temporal factors, including the time of day, day of the week, and season, impact electricity forecasting. Electricity demand fluctuates throughout the day due to varying consumer behaviors and activities, leading to peaks and troughs in demand during different periods. In addition, electricity demand patterns often differ on weekdays versus weekends. Weekdays typically see higher demand as commercial and industrial activities increase, while residential demand may be more dominant on weekends. Seasonal changes also affect both electricity demand and supply, with cooling demands increasing during summer and heating requirements rising during winter (Khatoon et al., 2014).

Calendar factors, such as holidays and special events, can impact electricity forecasting. On public holidays, electricity demand patterns often deviate from normal weekdays due to reduced industrial and commercial activities. Large-scale events like sports tournaments, concerts, or festivals can significantly affect electricity demand in the hosting regions (Khatoon et al., 2014).

Economic factors, such as GDP, GNP, and electricity prices, have a notable influence on load forecasting. Economic development, living standards, and price

elasticity of electricity consumption affect load consumption patterns. Economic factors play a prominent role in long-term load forecasting, while their impact on short-term forecasting is relatively minor (Khatoon et al., 2014).

User characteristics, such as customer type, size, and electrical equipment quantity, impact load forecasting. Different customer categories exhibit distinct load curves, including residential, commercial, and industrial consumers. Customer factors have a relatively lower influence on load forecasting, but the variations among different customer classes must be considered (Fahad & Arbab, 2014).

Random factors encompass sudden load variations or spikes caused by large industrial loads, agricultural demand, and special events. Sporting events, cultural celebrations, and wedding seasons also contribute to load forecast uncertainty. Random factors have a relatively smaller impact on load forecasting but may introduce unpredictability and volatility under specific circumstances (Khatoon et al., 2014).

Geographical variations and customer class distinctions can influence load curves. Load consumption in rural areas differs from urban areas, and load patterns may vary based on consumer categories. Other factors have a relatively minor impact on load forecasting, but accounting for geographical variations and consumer class differences enhances the accuracy and reliability of load predictions (Fahad & Arbab, 2014).

The consideration of these factors and understanding their degrees of influence improves the accuracy and reliability of load forecasting models, thereby supporting efficient power system planning and distribution.

Chapter 3.3 The current challenge in electricity forecasting

Electricity demand forecasting is a complex task, fraught with numerous challenges. These challenges range from extreme events and uncommon situations, the dynamic and stochastic nature of electricity demand, limited data availability for distribution grids, to the intermittency of renewable energy sources (Wada, Hori, & Taniguchi, 2020; Yadav, Jain, Sharma, & Bhakar, 2021; Ziekow, Doblander, Goebel, & Jacobsen, 2013; Chen, Gupta, & Tragoudas, 2022).

Chapter 3.3.1 Challenge in general

❖ Extreme Events and Uncommon Situations

Extreme events such as severe weather, natural disasters, and large-scale human behaviors significantly impact electricity demand. For instance, Watson, Spaulding, Koukoula, & Anagnostou (2022) found that extreme weather events can cause significant power outages. Similarly, the COVID-19 pandemic led to a significant reduction in electricity demand due to changes in human behavior (Xu, Gao, Li, & Qian, 2021).

❖ Dynamic and Stochastic Nature of Electricity Demand

The dynamic and stochastic nature of electricity demand also poses a challenge. Electricity demand exhibits time-dependent fluctuations, long-term economic trends, and stochastic variability (Chen & Majda, 2020). Furthermore, the limited availability of granular data at the distribution level makes it difficult to accurately forecast electricity demand (Akhtar et al., 2023).

❖ Limited Data Availability

Limited data availability is a common challenge in AI applications, and electricity forecasting is no exception. The lack of cleaned and structured data, especially when policies have changed, can hinder the performance of AI models. For instance, a change in energy policy might lead to a shift in electricity usage patterns, but if the data reflecting this change is not available, the AI model might fail to capture this shift.

❖ Intermittency of Renewable Energy Sources

The intermittency of renewable energy sources, such as wind and solar, poses another challenge. These sources are highly dependent on weather conditions, leading to fluctuations in their output. Balancing renewable supply with consumer demand is a critical task, and accurate forecasting is key to achieving this balance. The fluctuations in wind and solar generation make it difficult to balance renewable supply and consumer demand (Xu et al., 2021).

Chapter 3.3.2 Challenge in Japan

Firstly, as part of its decarbonization strategy, Japan is expanding solar, wind, and other renewables which made up 10% of its generation mix in 2021 (METI, 2021). The country's expansion of renewable energy has led to volatile renewable generation and uncertainties in load balancing and forecasting (Tanaka, 2013). Secondly, Japan imports over 90% of its fossil fuel supplies for electricity generation from foreign countries (Vivoda, 2016). As a result, dependence on imported fossil fuels means that electricity costs and availability are affected by global dynamics, which can have ripple effects on the economy and electric grid (Tanaka, 2013). For instance, the spot price of liquefied natural gas (LNG) in Asia

has surged nearly tenfold from average summer levels, resulting in crippling shortages in emerging countries with tight foreign exchange reserves, which in turn has caused electricity prices to also rise. (Hama, 2022). Such global shocks create ripple effects throughout Japan's economy and electric grid, increasing the complexity of forecasting nationwide electricity demand.

Furthermore, Japan's vulnerability to natural disasters, such as earthquakes and tsunamis, can cause significant damage to infrastructure and unpredictable load shifts (Tanaka, 2013). The country's seasonal climate also leads to extreme demands for electricity, especially during the hot summer months (Asahi, 2023). As people need to turn on their air-conditions in the hot weather, and heater in the cold winter.

Evolving social factors, such as demographics, urbanization, energy policies, remote working, automation, electric vehicles (EVs), and home solar, also impact electricity demand in Japan. For instance, the liberalization of Japan's electricity market and the transition to a low-carbon society have led to changes in electricity demand patterns (Tanaka, 2013).

In conclusion, electricity demand forecasting is a complex task that requires considering a wide range of factors. In Japan, these challenges are exacerbated by the country's unique energy landscape, vulnerability to natural disasters, and evolving social factors.

Chapter 3.4 The role of AI in electricity demand forecasting

AI plays a major role in maintaining the normal operation of power systems. This can be attributed to AI's powerful forecasting capabilities which can improve electricity forecasting accuracy and reduced forecast errors, efficient processing of

complex and high-dimensional data, automation and intelligent management of the forecasting process, and adaptation to changes in data patterns (Danish, 2023; Ahmad et al., 2022). AI's ability to process large volumes of data and identify complex patterns makes it particularly effective in improving the accuracy of electricity demand forecasts. This is especially important in the context of the dynamic and stochastic nature of electricity demand (Ngo et al., 2022).

In Japan's electricity market, AI plays a crucial role due to the country's unique energy landscape. The limited interconnections and variability of renewable sources make accurate demand forecasting particularly challenging. AI can help address these challenges and support Japan's goals of achieving zero carbon emissions and transitioning to a low-carbon energy system (Xu et al., 2019). AI also contributes significantly to the efficient management of Japan's power system. It aids in the efficient allocation of power resources, the development of reasonable pricing strategies, and the prediction of solar and wind power generation (Antonopoulos et al., 2020; Mhlanga, 2023).

Despite the benefits, the application of AI in electricity demand forecasting also presents challenges. These include the need for high-quality data, the complexity of AI models, and the need for continuous model updating and validation. However, with ongoing advancements in AI technologies and increasing digitalization of the power sector, these challenges can be effectively addressed (Danish, 2023).

Chapter 3.5 The need for human-AI collaboration in Japan electricity demand forecasting

In forecasting electricity demand in Japan, artificial intelligence and human collaboration are necessary. The power system is a complex system involving many variables and uncertainties. Although AI can efficiently process large-scale data, it is difficult to fully grasp the social factors affecting electricity demand, such as climate change, economic activities, and holidays. Therefore, human experts need to intervene, utilize their experience and expertise, help AI understand these influencing factors, and improve the accuracy of forecasting. In addition, the results of electricity demand forecasting will directly affect the dispatch and operation of the grid, involving a large number of economic and social benefits. Therefore, human experts need to review and judge the forecast results.

When implementing AI and human expert collaborative forecasting, steps can be taken to collect historical electricity demand data, select suitable AI models, develop human-machine collaboration mechanisms and processes, organize expert training, and establish authoritative expert review agencies. AI and human experts jointly discuss different forecasting models and solutions, analyze their advantages and disadvantages, and choose a hybrid model that takes into account both AI and expert knowledge based on this.

However, AI and human collaborative forecasting also face some challenges, such as contradictions between models and expert knowledge, model interpretability issues, subjectivity and limitations of expert knowledge, coordination issues, experts' acceptance of new technologies, and data and computing resource limitations. These issues can be solved by strengthening human-computer communication, improving model interpretability, enhancing

expert training and education, designing effective collaborative mechanisms, and adopting new technologies such as cloud computing.

In summary, AI and human collaborative forecasting of electricity demand is a challenging but opportunity-filled process. In this process, AI can help us process and analyze large-scale data and identify key factors affecting electricity demand. At the same time, human experts can provide industry knowledge and experience for AI models to help AI better understand and interpret complex electricity demand patterns. Only through deep human-machine collaboration can we achieve more accurate and comprehensive electricity demand forecasting, better meet the needs of Japan's power system, and promote its transformation into a smart grid.

In summary, forecasting electricity demand in Japan requires partnership between AI and human experts. While AI handles data analysis at scale, human judgment and knowledge fill critical gaps. Experts provide context on societal factors, evaluate AI models and results, and oversee system impacts. Collaboration eases challenges around conflicting insights, limited expertise, model transparency, and resource constraints. Discussion and education help address barriers, enabling more accurate, well-rounded forecasts.

A collaborative approach is key to managing the complexity of Japan’s power system and policy goals. AI and human experts together determine the most suitable forecasting methods, balancing strengths. Coordination through defined processes and governance maximizes the benefits of partnership. Progress relies on openness to new technologies and continued learning.

Overall, the analysis highlights the sophistication needed for electricity demand forecasting today and arguments for hybrid human-AI solutions. With sensitive, complex systems like energy at stake, single-method approaches risk missing key inputs or producing unreliable outcomes. Partnership is presented as the prudent

path forward, albeit requiring work to overcome obstacles. For Japan, human-AI collaboration seems essential to navigating additional layers of complexity and achieving an affordable, sustainable energy transition. The discussion reflects the thoughtful, forward-looking perspective needed for progress on issues like these.

Chapter 3.6 The case of extreme hot weather in Japan

This study utilized actual electricity demand data from Japan to conduct data analysis and visualization in order to show sudden fluctuations in demand associated with extreme weather events. as the chapter mentioned before, this a special events that can possibly not be predicted well by either statistical learning algorithm and machine learning algorithm. Identification of demand surges linked to extreme temperatures was accomplished through a three-step process:

- Identify extreme weather: extreme temperatures were defined statistically as those falling within the top and bottom 5th percentiles of the observed temperature distribution. The 5th and 95th percentiles were calculated from the temperature dataset.

- Identify sudden increase in electricity demand: sudden increases in electricity demand were operationalized as hourly demand changes exceeding one standard deviation above the mean change. Hourly differences in electricity usage were computed and instances exceeding the standard deviation threshold identified.

❖ Match extreme weather and sudden increase in electricity demand: cases were identified where an extreme temperature event was followed within 24 hours by a sudden increase in electricity demand. Each extreme temperature event was iteratively examined to determine if a demand spike occurred in the subsequent 24-hour period.

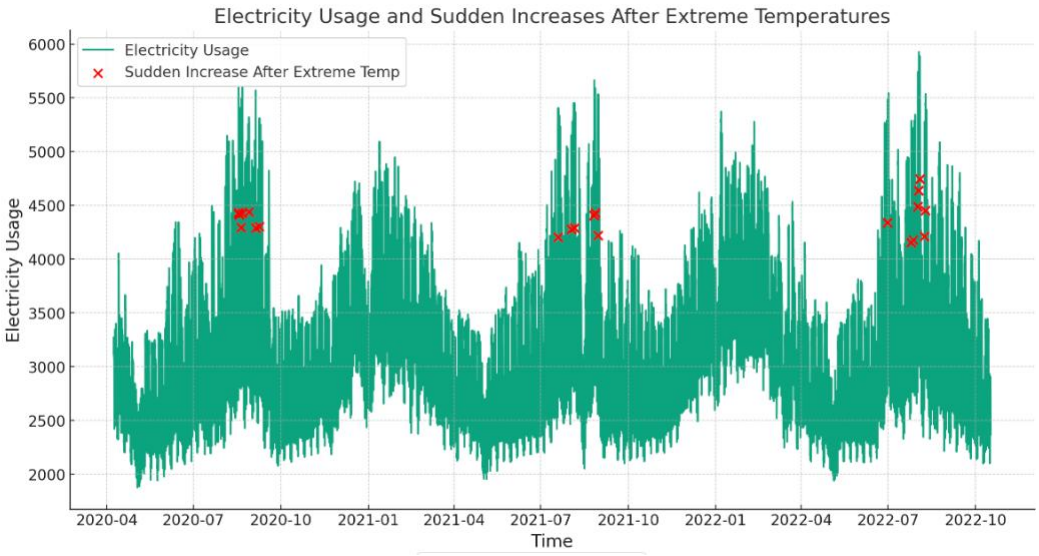

Figure 3.1 Electricity usage and sudden increases after extreme temperatures

As shown in figure 3.1, after the application of this three-stage analysis procedure yielded specific dates on which extreme temperatures were temporally linked to sudden demand surges in the data: August 17, 18, 20, 21, 28, September 4, 8, 2020; July 19, August 2, 5, 26, 27, 30 2021; June 30, July 25, 27, August 1, 2, 3, 8, 9 2022.

The concentration of these instances in summer months across three recent years suggests associations between extreme heat and spikes in electricity demand. These dates are spread across three years, and most of them occur in the summer

months (June, July, August, and September). This suggests that extreme temperatures and their impact on electricity demand could be associated with hot summer weather.

Through analyzing two specific dates from above, August 9, 2022, and August 26, 2021. The value of 1.98 standard deviations (std) is commonly used in statistics as a threshold for identifying outliers in a normal distribution. Approximately 95% of the data in a normal distribution falls within ±1.96 std from the mean, so values that exceed this range can be considered unusual or extreme. In this analysis, we have used a slightly more stringent threshold of ±1.98 std to highlight the most extreme fluctuations in electricity usage.

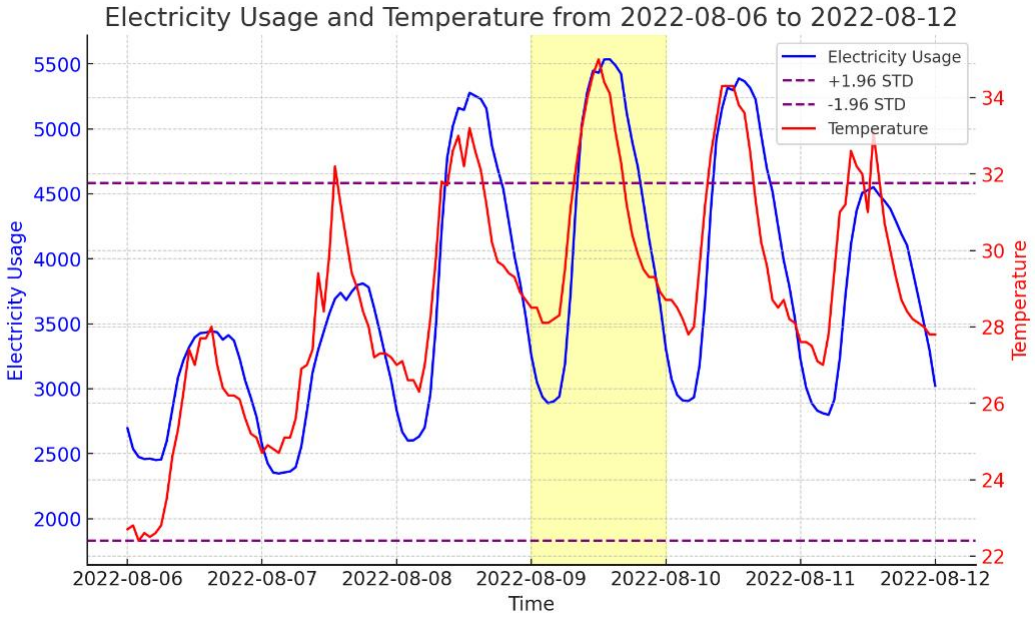

Figure 3.2 Electricity usage and temperature around August 9, 2022

In the dual-axis time series plots of figure 3.2 and figure 3.3, the purple lines represent these thresholds of ±1.98 std from the mean electricity usage. The blue

line represents the actual electricity usage data, and the red line shows the temperature data. The yellow shaded areas highlight two specific dates of interest: August 9, 2022, and August 26, 2021. Upon examining these plots, we can see that the electricity usage on both highlighted dates significantly exceeds the upper threshold of +1.98 std. This suggests that the electricity usage on these dates is unusually high compared to the typical usage pattern.

Both of these dates correspond to periods of extreme temperature, as shown by the peaks in the red line. This suggests a correlation between extreme weather conditions and unusual spikes in electricity demand, which is consistent with our understanding that hot weather can lead to increased use of air conditioning and other cooling devices, thus driving up electricity demand.

These findings highlight the impact that extreme weather can have on electricity demand, and the challenges it poses for electricity demand forecasting. Traditional forecasting models may not fully account for the effects of extreme weather, leading to underestimations of demand during such periods. This can result in insufficient electricity supply, leading to blackouts and other disruptions. Thus, incorporating weather data and identifying potential outliers in advance can be crucial for improving the accuracy of electricity demand forecasts and ensuring a reliable electricity supply.

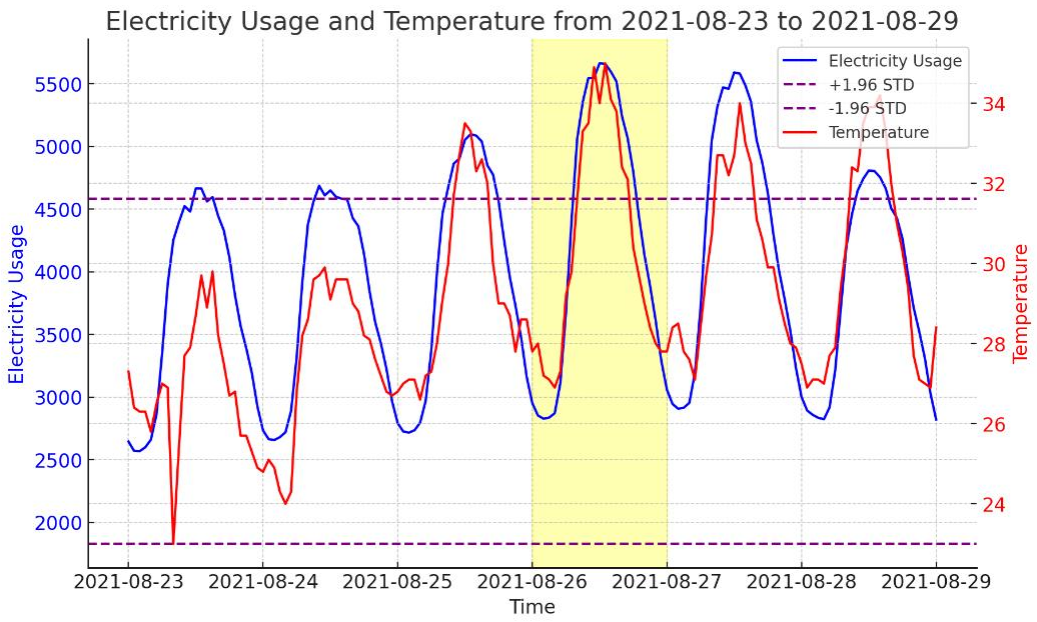

Figure 3.3 Electricity usage and temperature around August 26, 2021

Chapter 4 - Research Methodology

Chapter 4.1 Overview

This research employs a mixed-methods approach combining machine learning forecasting with human-computer interaction. The methodology chapters are organized as follows: Data collection and processing lays the foundation by obtaining quality datasets and preparing the time series data. Data analysis provides initial insights and relationships to guide model selection and training. The model selection chapter compares different forecasting models and justifies the choice of a flexible Generalized Additive Model (GAM). Rigorous steps for training the GAM are then outlined. Evaluation methods discusses techniques like k-fold cross validation and holdout sets to assess model accuracy, along with error metrics like RMSE and MAPE. This comprehensive methodology aims to produce an accurate and reliable electricity demand forecasting model. The trained model is then incorporated into an interactive visualization system to examine how human experts can interact with the model forecasts. This mixed-methods approach balances computational rigor in terms of data processing, model selection and training, with a human-centered perspective that allows domain experts to apply their insight. The overall methodology seeks to advance electricity demand forecasting through an integrated approach combining machine learning and human input.

Chapter 4.1 provides an overview of the research methodology which includes data collection, machine learning modeling and human interaction experiments. Chapter 4.2 discusses the data collection process involving weather and electricity demand data from reliable sources. The data is stored securely and handled rigorously. Chapter 4.3 performs a historical data analysis of the electricity and weather datasets to gain insights. Clear seasonal patterns are identified, showing higher demand in summer and winter. Electricity usage also varies based on time

of day and other factors. Chapter 4.4 discusses current electricity demand forecasting methods used in Japan, including Bayesian spatial models, scenario analysis and Ensemble Kalman Filter. Chapter 4.5 provides an overview of parametric, non-parametric and semi-parametric models, and their suitability for electricity demand forecasting. Chapter 4.6 introduces the General Additive Model (GAM), a type of semi-parametric model that can handle complex non-linear relationships. GAMs have been used effectively for electricity demand and price forecasting. Chapter 4.7 discusses different model types and criteria for selecting an appropriate model for electricity demand forecasting. Chapter 4.8 outlines the general model training process, and specifically the steps involved in training the GAM model. Feature selection, parameter tuning and avoiding overfitting/underfitting are important. Chapter 4.9 covers evaluation methods and error metrics like RMSE and MAPE that can be used to assess the performance of forecasting models. The choice of metric depends on the specific needs of the forecasting problem.

Chapter 4.2 Data collection

In this study, we primarily rely on data from two authoritative sources: Tokyo Electric Power Company (TEPCO) and the Japan Meteorological Agency. The openness and accuracy of the data from these institutions have been widely recognized, providing a solid foundation for the rigor and precision of our research. The collected data mainly covers the following aspects: hourly temperature, atmospheric pressure, relative humidity, and hourly electricity demand.

Category	Details
Data Sources	Tokyo Electric Power Company (TEPCO), Japan Meteorological Agency
Data Types	Hourly temperature, atmospheric pressure, relative humidity, precipitation, and electricity demand data represented as floating-point numbers
Data Volume	From April 1, 2020, to October 2022
Data Quality	No missing values, high completeness
Collection Method	Manual collection, ensuring no infringement of private or confidential information
Data Processing	Conversion of hourly datetime objects into distinct features such as hour, day, week, month, season, day of the week, and week of the month

Table 4.1 Table of data information

Hourly temperature data offers comprehensive environmental information, as temperature typically has a direct impact on people's living habits and electricity consumption. For instance, colder winter months may necessitate increased use of heating equipment, while hotter summer months may lead to greater air conditioning usage, thereby affecting electricity demand. Atmospheric pressure data serves as an indicator of basic weather conditions, with high pressure often associated with clear weather and low pressure potentially indicating overcast or

rainy weather. These varying conditions may influence living habits and electricity demand. Relative humidity data reflects weather comfort levels, with extremely high or low humidity potentially prompting individuals to adjust indoor environments, such as by using dehumidifiers or humidifiers, which in turn affect electricity demand. Hourly electricity demand data, expressed as floating-point numbers, reflects the power system's needs at specific times and provides a training target for our predictive model.

To ensure the reliability and scientific nature of our data processing, we store these data in local CSV files, a simple format that meets our computational requirements without introducing unnecessary complexity due to cumbersome data frameworks. Utilizing the efficient performance of the Apple M1 chip, resource usage during the computation process is highly economical. Throughout the data collection and storage process, we place significant emphasis on the originality and security of the data. We respect the value of each data point and refrain from making unnecessary modifications to the raw data. Furthermore, we store all data within a Google account associated with a University of Tokyo student account, ensuring data security and mitigating potential data leakage risks. This rigorous data handling approach lays a solid foundation for our subsequent research.

The above description of the data collection process serves as the basis for our subsequent data processing and model selection. In the next step, we will preprocess the data to better utilize it for subsequent analysis.

Chapter 4.2 Data processing

During the research process, the data preprocessing stage plays a crucial role, particularly when dealing with time series data that requires careful handling of its

characteristics. Our dataset, provided by Tokyo Electric Power Company (TEPCO) and the Japan Meteorological Agency, consists of floating-point time series data, including hourly temperature, atmospheric pressure, relative humidity, and electricity demand information. For these data, we conducted a series of standard preprocessing steps to prepare for further analysis.

First and foremost, by examining the properties of the dataset, the dataset was found to be complete, with no missing values. It was observed that both temperature and electricity demand exhibited prominent seasonal patterns. This indicates higher demand during winter and summer months for heating and cooling purposes, respectively, while demand remains lower in spring and autumn. To develop an accurate model, these seasonal factors need to be properly accounted for. Autocorrelation was also identified, indicating that electricity demand at a given time point is related to demand at adjacent time points. Additionally, a nonlinear relationship between temperature and electricity demand was discovered. These data characteristics are important considerations for model construction and forecasting.

Upon gaining a comprehensive understanding of the data's characteristics as shown in figure 4.1, we initiated the actual data processing work. Using Python's built-in Pandas library, we converted the raw date-time objects into more useful calendar information, including year, month, day of the week, hour of the day, weekday, week of the month, as well as indicators for holidays and weekends. Holiday information was obtained using the Holidays library, while weekend information was derived through simple function calls. This transformation is essential, as models tend to better understand and utilize structured information rather than raw date-time formats. Additionally, this approach facilitates subsequent feature engineering and model optimization.

temp	usage	year	month	day	hour	day_of_week	week_of_year
14.4	3124.0	2020	4	8	8	2	15
15.7	3274.0	2020	4	8	9	2	15
17.9	3266.0	2020	4	8	10	2	15
17.9	3260.0	2020	4	8	11	2	15
18.5	3106.0	2020	4	8	12	2	15

Figure 4.1 Data characteristics

Lastly, we employed the Pandas library to check data quality and consistency, ensuring the absence of duplicate values and data format errors. Although our data is complete and devoid of missing values, such verification remains necessary, as data quality and consistency directly impact the accuracy of subsequent analyses. In summary, our data preprocessing stage adheres to standard data science practices, ensuring that subsequent data analysis and model fitting are grounded in high-quality data. It is worth emphasizing that this is only the initial data processing step; during the actual data analysis and model fitting stages, we may need to conduct further data processing to accommodate specific model requirements or address unique issues within the data.

Chapter 4.3 Data analysis of electricity and weather datasets

Although performing historical data analysis is not the main task of this study, this section will present the data analysis based on the historical data, which is important for a better understanding of the datasets we are dealing with.

The processed dataset contains 22,117 rows and 3 columns spanning 2 years from April 8, 2020 to October 29, 2022 with temperature in Celsius and an

unspecified usage amount as the fields. The data was loaded directly from the CSV file without any preprocessing and initial exploration did not reveal any obvious quality issues, though more rigorous assessment may be needed. The data is complete with no missing values and the temperature has a mean of 17.75°C, standard deviation of 8.15, min of -3.1°C, and max of 36.7°C, while the usage amount has a mean of 3205.97, standard deviation of 702.18, min of 1877, and max of 5927.

Moreover, this figure 4.2 shows how average daily temperatures and electricity usage vary over time. As you can see from the graph, both temperature and electricity usage show a clear pattern of seasonal variation. Specifically, temperatures are higher in the summer and lower in the winter, while electricity usage is higher in the summer and winter and lower in the spring and fall. This may reflect an increase in electricity use during seasons of higher or lower temperatures, possibly due to the need for more electricity for cooling or heating.

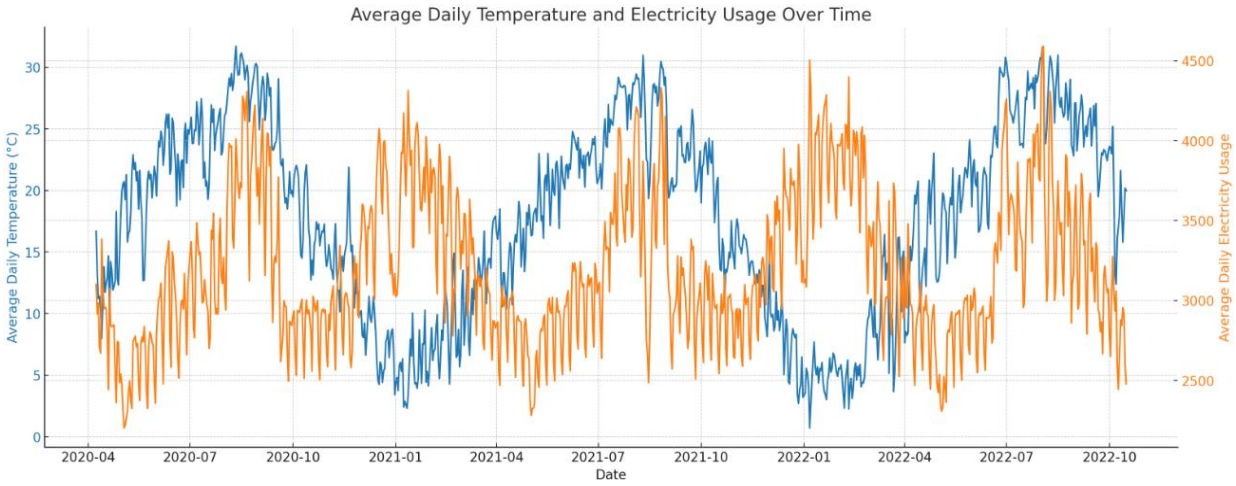

Figure 4.2 Average daily temperature and electricity usage

These two histograms in figure 4.3 show the distribution of temperature and electricity usage. From the histogram of the distribution of temperature, we can see that the distribution of temperature appears to be bimodal, possibly reflecting seasonal changes in temperature over the course of the year. From the histogram of the distribution of electricity usage, we can see that the distribution of electricity usage appears to be close to normal, but with some skewness on the side of high electricity usage.

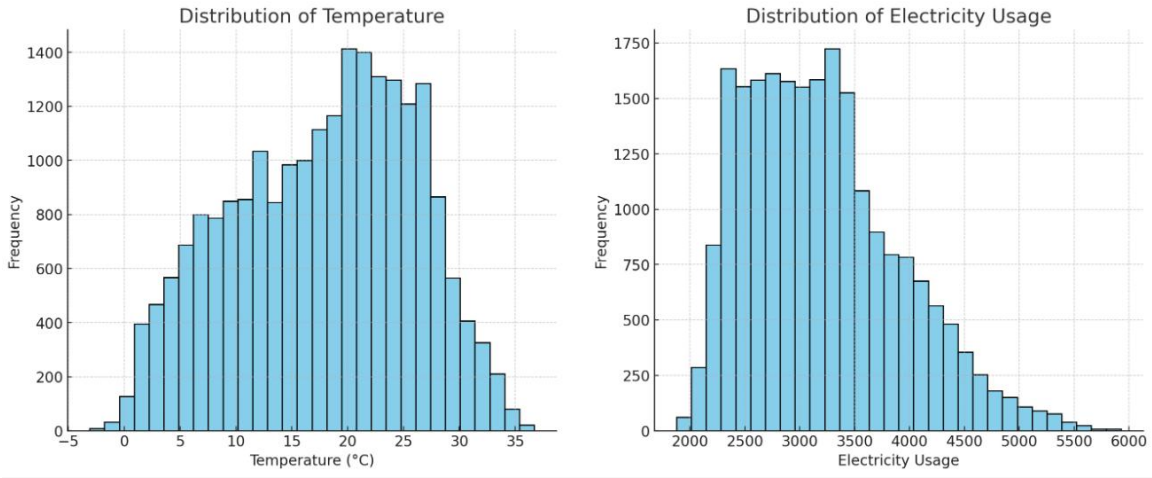

Figure 4.3 Distribution plots of temperature and electricity usage

These two heat maps in figure 4.4 show the distribution of average electricity usage and average temperature for different hours of the day and different months. The darker the color, the higher the electricity usage or temperature. From the heat map of electricity usage, electricity usage peaks in summer (June-August) and winter (December-February), while lower in spring (March-May) and autumn (September-November). This may be because in summer and winter, the use of air conditioners and heaters increases due to weather conditions (such as high or low temperatures), increasing the demand for electricity. In spring and autumn,

electricity demand is relatively low due to mild weather. From the heat map of temperature, we can see that temperature is higher during the day and in summer.

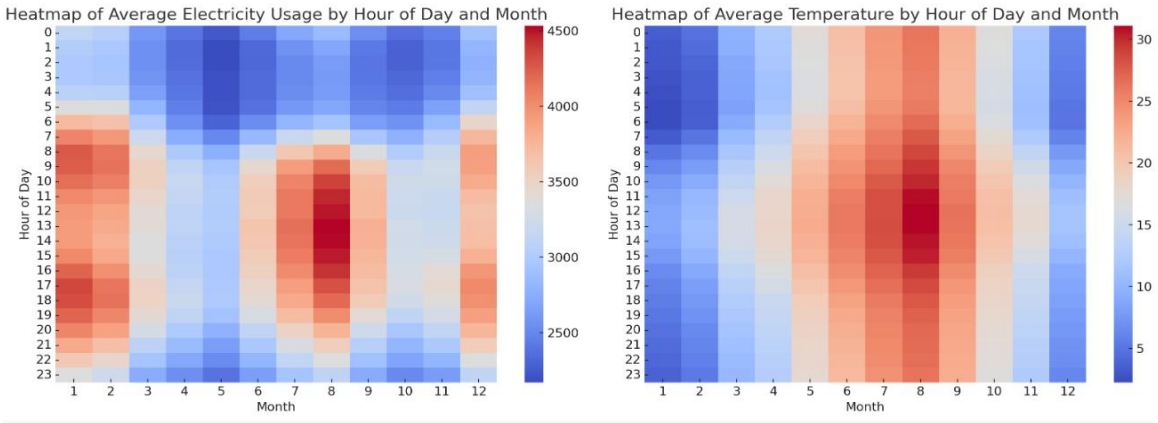

Figure 4.4 Heatmap of average electricity usage and temperature by hour of day and month

During the early hours (0-5 am), electricity usage is relatively low, probably because most people are asleep during this time period. Then around 7 am, electricity usage begins to rise, reaching the morning peak, probably because people start to get up and start their daily activities. Around 7 pm, electricity usage peaks, probably because people's activities at home increase after work, such as cooking, bathing, watching TV, etc. This pattern may reflect people's daily activity patterns and living habits.

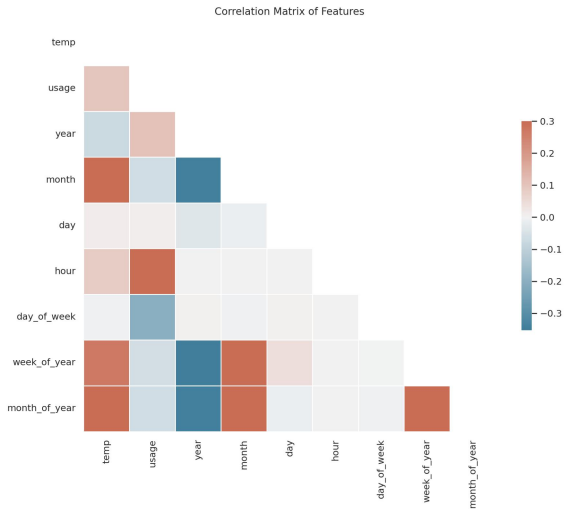

Figure 4.5 Correlation matrix of features

In this heat map in figure 4.5, darker areas indicate stronger correlation. We can see that there is a strong correlation between electricity usage (usage), hour (hour) and temperature (temp). This may be because electricity usage increases during certain hours of the day (such as morning and evening) and when the temperature is higher or lower.

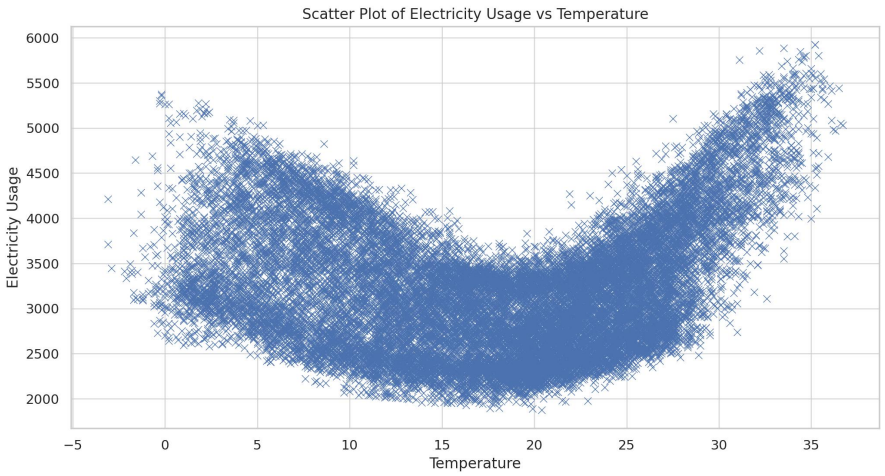

Figure 4.6 Scatter plot of electricity usage with temperature

It can be seen from the scatter plot in figure 4.6 that there is a "U-shaped" relationship between temperature and electricity usage. That is, when the temperature is lower or higher, electricity usage will increase, while when the temperature is in a moderate range, electricity usage is lower. This may be because when the temperature is lower, people need to use more electricity to heat, and when the temperature is higher, people need to use more electricity to cool. This pattern may reflect people's electricity usage behavior in the face of different weather conditions.

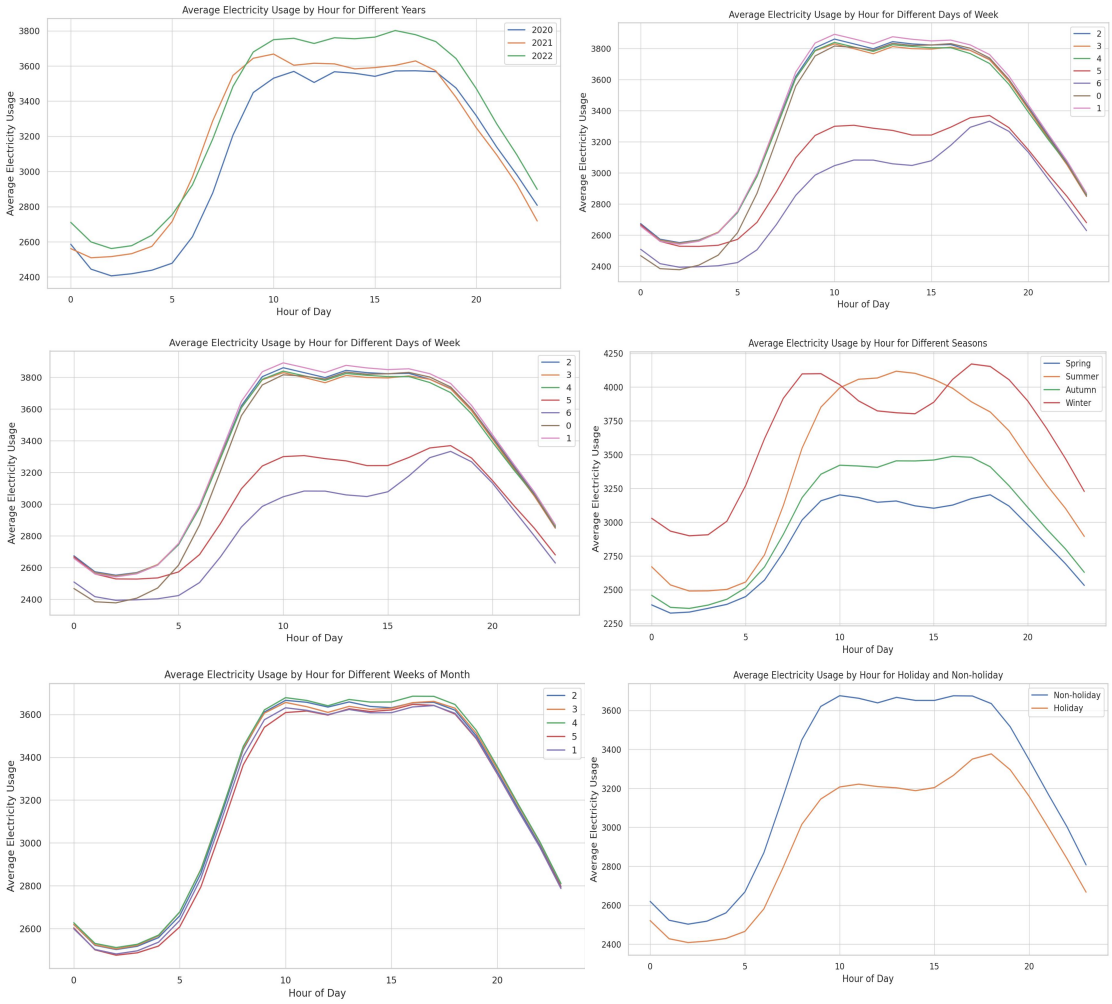

Figure 4.7 Average electricity usage by different datetime objects

In terms of finding patterns in different years and 24-hour electricity usage patterns. As shown in figure 4.7, the daily average electricity usage patterns throughout the day were roughly the same in different years, suggesting that electricity usage behavior is relatively stable in the short term. Different Days of the Week and 24-hour Electricity Usage Patterns: Electricity usage patterns on weekdays (0-4) and weekends (5-6) differed to some extent. This may be due to different activity patterns and living habits on weekdays and weekends. Four Seasons and 24-hour Electricity Usage Patterns: In different seasons, the daily average electricity usage patterns throughout the day showed significant differences. In particular, electricity usage during peak hours in the morning and evening was higher in summer and winter, possibly due to the need for more electricity for cooling and heating.Different Weeks of the Month and 24-hour Electricity Usage Patterns: In different weeks of the month, the daily average electricity usage patterns throughout the day were roughly the same, with no significant differences.

Holidays and Non-holidays' 24-hour Electricity Usage Patterns: Electricity usage patterns on holidays and non-holidays differed to some extent. This may be due to different activity patterns and living habits on holidays and non-holidays.

In conclusion, analysis of the historical electricity usage and temperature data reveals several insights. There are clear seasonal patterns, with usage peaking in summer and winter - likely due to increased heating and cooling needs. Usage also shows daily peaks in the morning and evening, reflecting human behavioral patterns. Temperature and usage have a U-shaped relationship, with higher usage at temperature extremes. While patterns are relatively stable annually and across different weeks, they differ on weekends vs. weekdays and holidays vs. non-

holidays due to shifts in human activities. Overall, the analysis illustrates how electricity usage is driven by seasonal, daily, and behavioral factors and suggests opportunities to improve load forecasting, planning, and management through deeper understanding of these relationships. Further rigorous statistical analysis would strengthen these findings and their applicability to forecasting models.

Chapter 4.4 Current electricity demand forecasting method in Japan

In the realm of electricity demand forecasting, Japan has adopted a variety of innovative and effective methodologies that cater to the unique characteristics of its electricity market. This paper aims to discuss the specifics of these methodologies, examining their advantages, disadvantages, and performance in the Japanese context.

One of the key methodologies employed in Japan for electricity demand forecasting is the Bayesian Spatial Autoregressive ARMA approach, which is particularly adept at handling spatial and temporal correlations inherent in electricity demand data (Hernandez, et al., 2014). The Bayesian component of this model allows for the integration of prior knowledge, thereby enhancing the accuracy of the forecasts. Despite the potential challenges in terms of computational efficiency and interpretability posed by the complexity of this model, it has proven to be effective in the Japanese context, given the country's intricate electricity grid and the importance of accurately forecasting demand to ensure grid stability (Hernandez, et al., 2014). Additionally, scenario analysis has become a pivotal method in Japan, particularly for long-term electricity supply and demand forecasting (Cap, 2022). This approach is valuable for accommodating various potential future scenarios, such as shifts in energy policy or technological advancements. While scenario analysis provides a broad perspective on potential

future developments, it may lack precision in short-term forecasts. Nevertheless, it remains a crucial tool for strategic planning in Japan's electricity sector (Cap, 2022).

Japan also employs the Ensemble Kalman Filter for electricity load forecasting and analysis (Takeda, Tamura & Sato, 2016). This method is well-suited to handling non-linear and non-Gaussian state-space models, which are common in electricity load data. Moreover, it allows for the incorporation of measurement errors into the model, thereby improving the accuracy of the forecasts. However, the Ensemble Kalman Filter requires a substantial amount of historical data for accurate forecasting, which may limit its applicability in certain situations. Despite this, its ability to handle complex data structures makes it a valuable tool in the Japanese electricity market (Takeda, Tamura & Sato, 2016).

In conclusion, the selection of these models for electricity demand forecasting in Japan reflects the complex nature of electricity demand in the country, which is influenced by a variety of factors and exhibits both spatial and temporal correlations. The effectiveness of these models in the Japanese context underscores the importance of tailoring forecasting methodologies to the specific characteristics of the electricity market.

Chapter 4.5 Parametric, non-parametric and semi-parametric model

In the field of electricity demand forecasting, various types of models have been employed, each with its unique characteristics and strengths. These models can be broadly categorized into parametric, non-parametric, and semi-parametric models. The choice of model type depends on the specific characteristics of the data and the forecasting requirements. This section provides an overview of these three

model types, their characteristics, use scenarios, and examples of forecasting algorithms that fall under each category.

- Parametric Models

Parametric models are a class of models that assume a specific functional form or distribution for the data. These models are characterized by a finite set of parameters, which are estimated from the data (Gautam & Singh, 2020). Examples of parametric models used in forecasting include linear regression models and autoregressive integrated moving average (ARIMA) models. Parametric models are particularly useful when the underlying distribution of the data is known or can be reasonably assumed. They are computationally efficient and provide interpretable results. However, their performance can be limited if the assumed distribution does not fit the data well (Mahmoud, 2021). In the context of electricity demand forecasting, parametric models such as ARIMA have been widely used. These models are capable of capturing linear relationships in the data and can handle trends and seasonality, which are common characteristics of electricity demand data (Weron & Misiorek, 2008).

- Non-parametric Models

Non-parametric models, on the other hand, do not make strong assumptions about the functional form or distribution of the data. These models provide more flexibility and can model complex, non-linear relationships. Examples of non-parametric models include decision trees and kernel regression models (Gautam & Singh, 2020). Non-parametric models are particularly useful when the distribution of the data is unknown or complex. They can model non-linear relationships and

interactions between variables, which can be beneficial in electricity demand forecasting where such complexities often exist. However, non-parametric models can be computationally intensive and may require large amounts of data to produce accurate forecasts (Härdle, 2004). In the realm of electricity demand forecasting, machine learning algorithms such as decision trees and random forests, which are non-parametric models, have been employed. These models can capture complex, non-linear relationships and interactions between variables, making them suitable for electricity demand forecasting (Gautam & Singh, 2020).

- Semi-parametric Models

Semi-parametric models are a hybrid of parametric and non-parametric models. They include both parametric and non-parametric components, allowing them to capture both linear and non-linear relationships in the data. Examples of semi-parametric models include generalized additive models (GAMs) and semi-parametric regression models (Fan & Hyndman, 2012). Semi-parametric models offer a balance between the interpretability of parametric models and the flexibility of non-parametric models. They can handle both linear and non-linear relationships and can model interactions between variables. However, they can be more complex and computationally intensive than purely parametric or non-parametric models (Liu et al., 2006). In electricity demand forecasting, semi-parametric models such as GAMs have been used. These models can capture the linear trend and seasonality in the data, as well as non-linear relationships and interactions between variables. This makes them particularly suitable for electricity demand forecasting, where such complexities often exist (Fan & Hyndman, 2012; Krstonijević, 2022).

In conclusion, parametric, non-parametric, and semi-parametric models each offer unique strengths in electricity demand forecasting. The choice of model type should be guided by the specific characteristics of the data and the forecasting requirements. Future research could explore the development and application of hybrid models that combine the strengths of these different model types to further improve forecasting accuracy.

Chapter 4.6 Introduction of General Additive Model (GAM)

Generalized Additive Models (GAMs) are a type of statistical model that can be used to describe the relationship between a response variable and one or more predictor variables. They are an extension of Generalized Linear Models (GLMs) and allow for more flexibility in the relationship between the response and predictor variables by using smooth functions of the predictors (Wood, 2006).

$$g(E(Y)) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Equation 4.1 The formula of General Additive Model (GAM)

Where:

- Y : is the response variable.
- $X_1, X_2, \ldots, X_p$: are the predictor variables.
- g : is the link function
- $E(Y)$: is the expected value of the response variable
- α: is the intercept

GAMs are particularly useful when the relationship between the response and predictor variables is not linear or when the relationship is not well understood. They can handle non-linear relationships and interactions between predictors without having to specify these relationships or interactions explicitly (Wood, 2006).

The advantages of the GAM model include its flexibility, its ability to handle different types of data and relationships, and its interpretability. However, it also has some disadvantages. For example, it can be computationally intensive, especially for large datasets, and it requires careful selection of the smoothing functions (Wood, 2006).

In the context of time series forecasting, the GAM model has been used effectively in the electricity market. The model's flexibility allows it to capture the complex patterns and trends in electricity demand and price data, leading to accurate forecasts (Hastie & Tibshirani, 1990). For example, Fan and Hyndman (2010) used a semi-parametric additive model to forecast short-term electricity demand and found that it outperformed other models in terms of accuracy. Similarly, Krstonijević (2022) used a GAM with automatic variable selection to forecast load and found that it provided accurate and reliable forecasts.

The performance of the GAM model in the electricity market can be attributed to its ability to capture the non-linear relationships and complex patterns in the data. For example, electricity demand and prices are influenced by a variety of factors, including weather conditions, time of day, and day of the week. The GAM model can accommodate these factors and their interactions, leading to accurate forecasts (Amato et al., 2021).

In addition to its use in forecasting, the GAM model has also been used in other areas of the electricity market. For example, Meier et al. (2019) used a GAM to

forecast short-term electricity prices and found that it provided accurate and reliable forecasts . In a subsequent study, Meier et al. (2020) compared the performance of a GAM with a deep artificial neural network (ANN) for short-term electricity price forecasting and found that the two models performed similarly, with the GAM providing slightly more accurate forecasts.

In conclusion, the GAM model is a flexible and powerful tool for forecasting. It can handle different types of data and relationships, making it suitable for a wide range of scenarios. In the context of the electricity market, the GAM model has been used effectively to forecast demand and prices, with studies showing that it outperforms other models in terms of accuracy. However, it also has some disadvantages, including computational intensity and the need for careful selection of the smoothing functions. Despite these challenges, the GAM model remains a valuable tool for forecasting in the electricity market.

Chapter 4.7 Model selection

Electricity demand forecasting is an essential aspect of power system management, and various methodologies have been employed to tackle this challenge effectively. Each method boasts unique strengths and limitations, necessitating a comprehensive understanding of these models and their applicability in the context of electricity demand prediction. This paper will discuss the major categories of forecasting models, highlighting their characteristics and suitability for electricity demand forecasting.

Artificial Neural Networks (ANNs) have recently gained prominence due to their exceptional ability to model complex non-linear relationships, a characteristic commonly found in electricity demand data (Román-Portabales et al., 2021).

Despite their strengths, ANNs have certain drawbacks, such as their black-box nature and the requirement for substantial computational resources. In contrast to ANNs, time series models like AutoRegressive Integrated Moving Average (ARIMA) are rooted in statistical theory and excel at capturing linear temporal dependencies, although they struggle with non-linear patterns (Hajirahimi & Khashei, 2019). To address this limitation, time series models are often combined with machine learning techniques, resulting in a synergistic effect that enhances overall forecasting accuracy. Machine learning algorithms, such as Support Vector Machines (SVMs) and Random Forests, are also increasingly employed alongside traditional statistical methods to improve forecasting accuracy (Grandon et al., 2023). These algorithms can handle high-dimensional data and capture complex non-linear relationships, but their performance is heavily reliant on the quality and quantity of the training data.

Hybrid models, which amalgamate different forecasting methods, have emerged as a promising solution for achieving higher accuracy (Hajirahimi & Khashei, 2019). These models are classified into three primary structures: parallel, series, and parallel-series. The parallel-series hybrid structure, which capitalizes on multiple models' strengths while mitigating their individual weaknesses, has been found to yield more accurate results than its counterparts. However, the complexity of these models could lead to challenges in interpretation and implementation. Grey models, designed to handle uncertain and incomplete information, have also been employed for electricity demand forecasting, particularly when data quality is low or scarce (Hajirahimi & Khashei, 2019). These models are often used in combination with other methods in a hybrid approach, but their performance may be limited in situations with abundant, high-quality data.

In summary, a multitude of models has been utilized for electricity demand forecasting, with hybrid approaches often employed to leverage the strengths of several methodologies. The choice of model depends on the specific characteristics of the electricity demand data and forecasting requirements, emphasizing the importance of a nuanced understanding of both the data and the models. Future research in this field will likely continue to explore and refine these methodologies, focusing on hybrid models and machine learning algorithms.

Chapter 4.8 Model training

Model training is a crucial process in machine learning.As shown in figure 4.8, It involves a series of steps designed to ensure that the model performs well not only on the training data but also on new, unseen data. These steps include: (1) splitting the data into a training set and a testing set, (2) feature selection or feature extraction, (3) training the model and optimising the parameters, and (4) making predictions and comparing them with the validation dataset (Sarker, 2021).

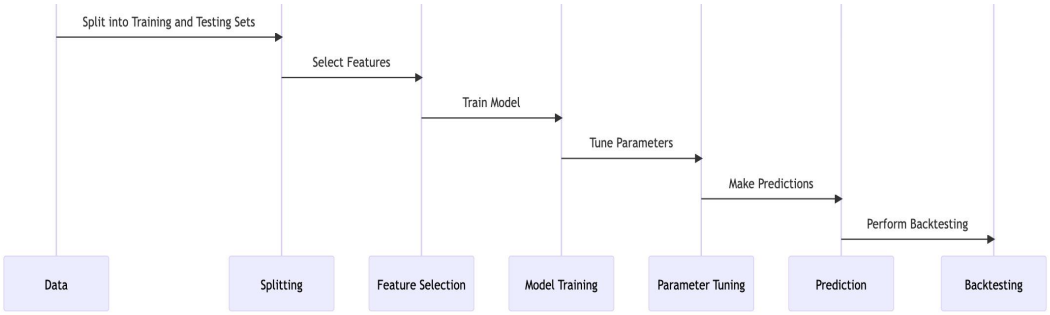

Figure 4.8 General model training process

In our study, we employ the Generalized Additive Model (GAM). The figure 4.9, shows the steps involved in the model training process. The process begins with data preparation, which includes collecting, cleaning, and engineering features

from the data. The next step is model selection, where the modeler chooses a model that is appropriate for the data. The model is then trained on the data, and its parameters are adjusted until it learns to make accurate predictions. Finally, the model is evaluated to see how well it performs. The flowchart also includes a feedback loop, which allows the model to be updated if it is not performing well.

The first step in our model training process is to split the data into a training set and a testing set at a ratio of 8:2. This ratio is chosen to ensure that the model has ample data for training while reserving a portion of the data to evaluate the model's performance. The division of data into separate training and testing sets is a common practice in machine learning, as it helps prevent overfitting, where the model performs well on the training data but poorly on new, unseen data (Ying, 2019).

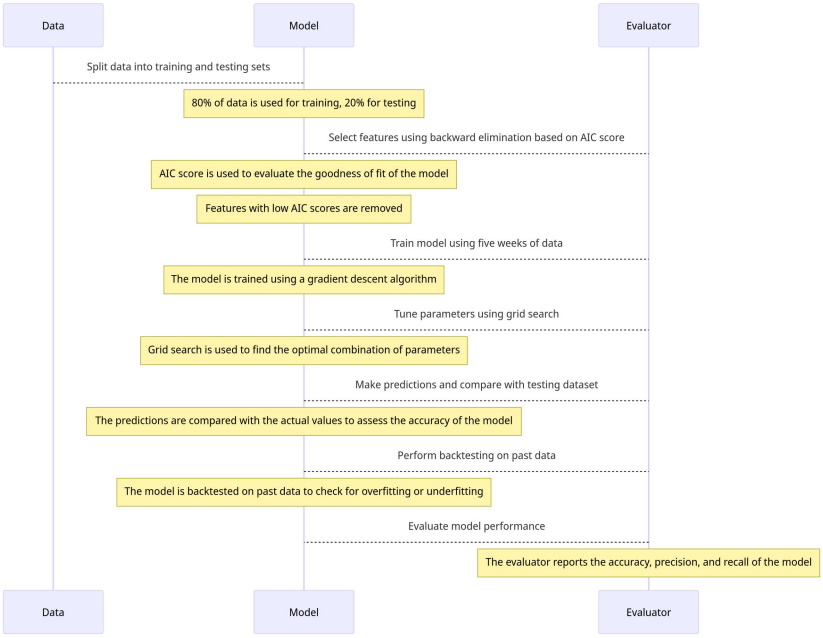

Figure 4.9 Flowchart the model processing process in the research

The next step in our process is feature selection. Not all features are useful for model prediction; some may introduce noise and negatively impact the model's

performance. Therefore, we use a method known as backward elimination based on the Akaike Information Criterion (AIC) score for feature selection (Aggrawal & Pal, 2021). The process of backward elimination involves initially training the model with all features and then gradually removing the least impactful features until the model's performance begins to significantly deteriorate. We choose AIC as our evaluation criterion because it considers not only the goodness of fit of the model but also the complexity of the model, effectively preventing overfitting (Bozdogan, 1987).

Once we have selected the most relevant features, we train our model using five weeks of data, with one week of data reserved for testing. During the parameter tuning phase, we utilise the grid search method built into pygam to optimise the parameters (Tutz & Binder, 2004). Grid search is an exhaustive search method that combines pre-set parameter values to find the combination that optimises model performance. Although this method is computationally intensive, it ensures that we find the global optimum.

After training the model and optimising the parameters, we use the model to make predictions and compare these predictions with the testing dataset. We also perform backtesting on past data to ensure that there are no issues with overfitting or underfitting. Overfitting occurs when a model performs well on the training set but poorly on the testing set, usually because the model is too complex and has learned the noise in the training set. Underfitting, on the other hand, occurs when a model performs poorly on both the training and testing sets, typically because the model is too simple to capture the complex relationships in the data (Sehra, Flores, & Montanez, 2021). To avoid both overfitting and underfitting, we need to strike a balance between model complexity and model generalisation ability.

In conclusion, the model training process is a delicate balance of selecting the right features, tuning parameters, and ensuring the model is neither too complex (leading to overfitting) nor too simple (leading to underfitting). Despite the complexity of this process, it is essential for building robust and reliable predictive models. However, it is important to note that no matter how complex or accurate our model is, errors are inevitable. Therefore, we must continually evaluate and adjust our model to improve its future predictions.

Chapter 4.9 Evaluation methods

Various techniques exist for evaluating the performance of time series forecasting models. Cross-validation methods like K-fold cross-validation systematically estimate model performance but may require modifications for time series data. Out-of-sample approaches using a holdout set are better suited for time series and provide realistic estimates of model performance. Prequential evaluation allows for continuous monitoring of model performance in online settings, though it is not suitable for all time series tasks. Choosing suitable evaluation methods based on the characteristics of the time series data and objectives of the evaluation is key. Different error measures like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) offer various perspectives on model accuracy. MAE and RMSE measure absolute errors while MAPE provides interpretable percentage errors. Selecting appropriate error metrics based on the specifics of the forecasting problem is essential for properly evaluating time series models.

Chapter 4.9.1 Overview of general evaluation methods

Various evaluation methods are employed to assess the accuracy and performance of time series forecasting models. The choice of evaluation method depends on the specific characteristics of the time series data and the objectives of the evaluation (Hyndman & Athanasopoulos, 2018). Cross-validation approaches, such as K-fold cross-validation, blocked cross-validation, and cross-validation with holdout, systematically estimate model performance by considering the temporal order of observations (Bergmeir & Benítez, 2012). However, they may not be directly applicable to non-i.i.d. data, such as time series, without modifications (Hyndman & Athanasopoulos, 2021). Out-of-sample (holdout) approaches are more suitable for time series data and can provide realistic estimates of model performance (Tashman, 2000). Prequential evaluation, which evaluates the model's performance in an online learning setting, allows for continuous model updates and performance monitoring in real-time scenarios, although it may not be suitable for all types of time series tasks (Dawid, 1984).

Chapter 4.9.2 Overview of measurement of forecast error

Different error measures can be used to evaluate the performance of time series forecasting models. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are commonly employed as they provide various perspectives on the accuracy of the models (Cerqueira et al., 2020; Dama & Sinoquet, 2021). MAE and RMSE are absolute error measures, with RMSE penalizing large errors more heavily than MAE. MAPE, on the other hand, provides a percentage error, which is easier to interpret in practical terms. It is essential to understand the properties of these measures when using them to

evaluate forecasting models and choose appropriate error measures based on the specific characteristics and requirements of the forecasting problem (Davydenko & Fildes, 2016; Shcherbakov et al., 2013).

Chapter 4.9.3 Evaluation methods of GAM model: an overview

In the realm of electricity demand forecasting, various error metrics are employed to evaluate the performance of different models. Each metric provides a unique perspective on the accuracy of the model, and the choice of metric often depends on the specific characteristics of the forecasting problem.

In the study by Gautam and Singh (2020), the Mean Absolute Percentage Error (MAPE) was used as the forecast error metric. This metric is particularly useful when dealing with time series data, as it provides a percentage error, making it easy to interpret across different scales. The authors found that non-parametric models generally outperformed parametric models in terms of MAPE, indicating their superior forecasting accuracy (Gautam & Singh, 2020). Fan and Hyndman (2012) used the Root Mean Squared Error (RMSE) to evaluate their semi-parametric additive model for short-term load forecasting. RMSE is a commonly used metric in regression analysis and machine learning, as it provides a measure of the average magnitude of the prediction error. The authors found that their model achieved a lower RMSE compared to other models, indicating its superior performance (Fan & Hyndman, 2012). Li et al. (2019) also used RMSE to evaluate their weekend load forecasting model based on semi-parametric regression analysis. The authors found that their model achieved a lower RMSE compared to other models, indicating its superior performance (Li et al., 2019). Weron and Misiorek (2008) used both the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE) to evaluate their spot electricity price forecasting models.

These metrics provide a measure of the average magnitude of the errors in a set of predictions, without considering their direction. The authors found that semi-parametric models generally outperformed parametric models in terms of both MAE and MAPE (Weron & Misiorek, 2008).

In conclusion, the choice of forecast error metric depends on the specific characteristics of the forecasting problem and the type of model being evaluated. However, regardless of the metric used, the goal is always to minimize the forecast error and improve the accuracy of the model.

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values observed. It is the square root of the average of squared differences between prediction and actual observation. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Equation 4.2 Formula of Root Mean Square Error (RMSE)

Where:

- ✓ y_i: is the actual value
- ✓ $\hat{y}_i$: is the predicted value
- ✓ n: is the number of observations

RMSE is a good measure of accuracy, but one of its problems is that it gives higher weight to larger errors. This means the RMSE should be more useful when large errors are particularly undesirable (Chicco, Warrens & Jurman, 2020).

Mean Absolute Percentage Error (MAPE) is a statistical measure to define the accuracy of a machine learning algorithm or a forecasting method in statistics. The formula for MAPE is:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Equation 4.3 Formula of Mean Absolute Percentage Error (MAPE)

Where:

- ✓ y_i: is the actual value
- ✓ $\hat{y}_i$: is the predicted value
- ✓ n: is the number of observations

MAPE has the advantage of being easy to understand and calculate. However, it has the disadvantage of being infinite or undefined if there are zero values in the data, which can happen if you are tracking residuals or differences (Chai, 2020).

In the context of electricity demand forecasting, both RMSE and MAPE can be useful. In our dataset, we do not have large errors in general, since I assume the forecasting algorithm has the ability to make appropriate forecasts. MAPE, on the other hand, could be another good choice if we want an intuitive measure of average error. However, this research does not need to compare the forecasting

results across different scales (e.g., across different regions with different average levels of electricity demand). Therefore, RMSE is the main forecast error measurement that has been applied in this research.

In conclusion, this chapter has presented an overview of the research methodology for electricity demand forecasting. Various models have been discussed along with their suitability for this application, highlighting the flexibility and accuracy of semi-parametric models like the GAM. The importance of a rigorous data collection and processing approach has been emphasized to ensure high quality inputs for the models. Careful model training, parameter tuning and evaluation are essential to develop robust forecasting models that avoid overfitting while achieving a good balance of complexity and accuracy. With a sound methodology in place, featuring suitable models, high quality data and a systematic process, accurate electricity demand forecasts can be generated to inform effective planning and management within the power system.

Chapter 5 - Human collaboration

Chapter 5.1 Overview

This chapter explores integrating human judgment into electric power demand forecasting through a collaborative framework between experts and AI systems. Judgmental forecasting leverages intuitive expertise and qualitative insights from human experts that statistical models struggle to incorporate. When historical data is limited or upcoming events are expected to disrupt patterns, judgmental forecasting allows experts' contextual knowledge to generate initial forecasts and adjust model outputs.

However, judgment is also prone to cognitive biases that reduce accuracy. Therefore, it is critical to structure the collaboration process to maximize the benefits of human expertise while minimizing bias. The interactive interface designed enables experts to thoroughly understand the prediction task context and machine outputs before making focused adjustments. Experts can choose to modify either locally or globally based on their domain knowledge. Confirming changes before saving prompts deliberation, while subsequent user surveys provide feedback to improve the process.

By respecting both human judgment and algorithms, this human-AI collaborative framework aims to produce more accurate, robust, and adaptive demand forecasts. The system provides necessary information and freedom for human experts to contribute their unique strengths while using AI to address systematic biases. The combination of complementary human and machine capabilities contributes to an overall forecasting process that is greater than the sum of its parts.

This chapter is organized into five main sections. First, an introduction to judgmental forecasting explains how human expertise can complement statistical models. Next, the benefits and risks of judgmental forecasting are discussed. The

third section outlines different types of judgmental forecasting approaches. Design principles for effective human-AI collaboration are then presented, highlighting how AI systems can support and enhance human judgment. Finally, details of the experimental design are provided, including the interactive interface, human participation, and user feedback mechanisms. The flow of the chapter highlights key aspects of integrating human factors into the forecasting process through a collaborative framework between experts and AI.

Chapter 5.2 Integration of human factor in forecasting

This section introduces judgmental forecasting as a method that relies on human judgment, intuition, and qualitative factors to make predictions. It highlights how judgmental forecasting differs from purely statistical models by incorporating expert opinions and contextual knowledge. Qualitative factors such as promotions, price changes, and competitor actions, which statistical models may struggle to consider, are taken into account. Judgmental forecasting is particularly useful when historical data is limited, allowing experts to rely on their knowledge and industry experience to generate initial forecasts.

Chapter 5.2.1 What is judgmental forecasting

Judgmental forecasting is a forecasting method that leverages human judgment, intuition, and qualitative factors to generate predictions (Lawrence et al., 2006). It stands in contrast to purely statistical models, incorporating expert opinions and contextual knowledge that domain experts possess due to their experience, situational awareness, and product familiarity (Edmundson et al., 1988; Lawrence et al., 2000). It accounts for qualitative factors like upcoming promotions, price changes, and competitor actions that statistical models typically do not consider due to difficulties incorporating such factors (Goodwin, 2000). Furthermore, it is particularly useful when historical data is limited, allowing experts to generate initial forecasts based on their knowledge and industry experience (Goodwin et al., 2014; Fildes et al., 2009).

Chapter 5.2.2 When Do We Use Judgmental Forecasting and Its Advantages

Judgmental forecasting is commonly used in supply chain and sales forecasting where qualitative factors and contextual knowledge play an important role (Lawrence et al., 2000). It is especially valuable in situations where experts can visually perceive trends and patterns in time series data (Lawrence et al., 1985; Lawrence & Makridakis, 1989). For new products with no historical sales data, experts can generate reasonable forecasts based on their knowledge, industry experience, and intuition (Goodwin et al., 2014; Fildes et al., 2009). Experts can account for upcoming events like promotions, price changes, and competitor actions in their forecasts (Edmundson et al., 1988). Judgmental adjustments of statistical forecasts can improve accuracy by incorporating relevant contextual information omitted by the models (Mathews & Diamantopoulos, 1986). Combining judgmental and statistical forecasts by averaging or weighting can produce more accurate forecasts by incorporating independent and complementary sources of information (Clemen, 1989; Fildes & Goodwin, 2007).

Chapter 5.2.3 Risks and Challenges of Judgmental Forecasting

While judgmental forecasts can be accurate, they are also prone to cognitive biases and systematic errors that reduce accuracy. Experts display cognitive biases like the representativeness heuristic and anchoring, where they assign too much weight to recent data and perceived patterns that are actually random noise (Hogarth & Makridakis, 1981). They also tend to overestimate trends and damp trends excessively, particularly downward trends (O'Connor et al., 1993). Overconfidence in their judgments and forecasts due to unrealistic optimism and illusion of control is another common issue (Kahneman & Riepe, 1998). Organizational and political pressures can also affect experts' judgments, leading to

biased forecasts (Oliva & Watson, 2009). Furthermore, judgmental adjustments of statistical forecasts frequently reduce accuracy when experts misinterpret noise as signal, make unnecessary adjustments, or overweight their adjustments (Willemain, 1991; Goodwin, 2000).

Chapter 5.2.4 Types of Judgmental Forecasting Methods

There are three key types of judgmental forecasting:

- Pure Judgmental Forecasts: Experts generate forecasts entirely based on their judgment and opinions, without referencing any statistical model. They leverage their contextual knowledge, industry experience, and intuition to make predictions (Fildes & Goodwin, 2007; Goodwin, 2000).
- Judgmental Adjustments of Statistical Forecasts: Experts adjust baseline statistical forecasts to incorporate relevant contextual information that the models omit. The process involves determining whether adjustments are needed and estimating the magnitude and direction of adjustments (Mathews & Diamantopoulos, 1986; Lawrence et al., 2006). Adjustments are warranted when the model is misspecified, new information arises, or errors are systematic (Sanders & Ritzman, 1992). The value of domain knowledge encapsulated in judgmental adjustments can significantly improve accuracy, especially when adjustments are large and negative (Mathews & Diamantopoulos, 1990; Fildes et al., 2009). However, many unnecessary adjustments reduce accuracy (Willemain, 1991; Goodwin, 2000).
- Combining Judgmental & Statistical Forecasts: This involves combining statistical forecasts with judgmental inputs through averaging or weighting to produce forecasts that leverage complementary strengths. Forecasts from

independent sources draw upon different information, thereby increasing the total information incorporated (Sanders, 1992; Clemen, 1989). Mechanical combinations often outperform judgmental combinations due to reduced bias (Goodwin & Wright, 1994). Combining a judgmental forecast with a statistical forecast tends to produce more accurate forecasts than combining two statistical forecasts (Lawrence et al., 1986). The optimal combination depends on the credibility of inputs and their relative strengths (Abramson & Clemen, 1995).

Chapter 5.2.5 Designing Effective Human-AI Collaboration in Judgmental Forecasting

In the era of artificial intelligence (AI), judgmental forecasting is evolving into a collaborative process between humans and machines. AI systems can provide optimal guidance, feedback, and restrictiveness to maximize the contributions of human judgment (Adya & Lusk, 2012; Fildes et al.,2006). A continuous feedback loop between the human experts and the AI system allows both to learn from each other, with the AI system identifying systematic patterns in the expert judgments and the experts providing qualitative insights and hard-to-model factors to improve the statistical models. Furthermore, the causal information identified by human experts plays a crucial role within this feedback loop. By representing these causal factors within its statistical models, the AI system can generate increasingly accurate baseline forecasts that require fewer adjustments, while still benefiting from the valuable qualitative insights and hard-to-model factors experts provide (Lim & O'Connor, 1995; Seifert et al., 2015).

Chapter 5.3 Introduction of the design of the experiment

Chapter 5.3 provides an introduction to the design of the experiment conducted in the study. This subsection begins by outlining the objectives and organization of the chapter, which focuses on designing an effective human-AI collaborative model to improve the accuracy, robustness, and adaptability of electric power demand forecasting. The chapter is organized into several subsections, including an overview of the design, the design of the user interface, the implementation of the experiment, and the evaluation of the results. Each subsection explores different aspects of the experiment design, highlighting the interactive interface, human participation, and user feedback. The chapter aims to showcase the collaborative process between humans and AI, emphasizing the integration of their unique strengths to achieve more accurate and reliable predictions.

Chapter 5.3.1 Objective and design

This study aims to design and implement an effective human-AI collaborative model to improve the accuracy, robustness and adaptability of electric power demand forecasting. The basis of this design is to recognize and respect the unique advantages of humans and AI, and to combine the strengths of both to achieve the goals. Specifically, generalized additive models (GAMs) are first used to model historical data and generate initial predictions. The results are then translated into visual forms to provide users with detailed and understandable technical information.

Stage	Description	Key Features
1: Interactive Interface	Develop an environment where users can easily access and understand the information.	Information of data analysis, Contextual information, Modeling information, Historical Data visualization, Comparison tool of different date
2: Human Participation	Respect and make full use of human subjectivity. Users are allowed to directly adjust the AI's prediction results.	Global modification, Local modification
3: User Feedback	Implement specific functions to improve the user experience.	Feedback function, Undo function, Reset function

Table 5.1 Introduction of three stages in interacting with user interface

As shown in table 5.1, the human-AI collaboration has three stages in our experiment. The first stage of the design focuses on developing the interactive interface. The core goal is to create an environment where users can easily access and understand the information. Therefore, the information presented to users is divided into two categories: technical information and background information. The technical information includes data analysis and modeling information. At this stage, users are allowed to select specific time periods for visualizing historical data to help them understand the basis of predictions in depth. An interactive tool is also designed for users to select two specific dates for comparison, and a 24-hour visualization of power demand and weather data for the comparison is presented to help users intuitively understand the changes in data. At the same time, non-quantitative information related to the prediction task is provided in the background information. This information is presented in text to help users understand the specific context and conditions of the prediction task from the context.

The second stage of the design is the human participation stage, which respects and makes full use of human subjectivity. At this stage, users are allowed to directly adjust the AI's prediction results. Users can make adjustments based on their expertise, experience or intuition, which is an essential part of the design because humans can provide additional information that the AI cannot obtain. Two modification methods are provided: global modification and local modification. Global modification allows users to fine-tune the overall prediction results with one-time up or down shifts for all data. Local modification allows users to adjust specific parts of the prediction results. This is because it is recognized that users may have a deeper understanding and expectation of specific time periods or situations. Therefore, a function is provided to allow users to modify the prediction results based on their own understanding.

In the final stage of the design, specific functions are implemented to improve the user experience. First, a feedback function is provided so that participants can immediately see the impact of their modifications on the prediction results, which helps them better understand the impact of their decisions on the results. Second, an undo function is provided so that participants can cancel their last operation at any time to reduce possible errors. Finally, a reset function is provided, which means that participants can return to their initial state at any time. These functions are designed to provide a user-friendly environment so that participants can operate more freely and confidently to improve the overall prediction performance.

In summary, the design emphasizes the collaboration between AI and humans while respecting and bringing out human subjectivity. It is believed that by combining the strengths of AI and humans, higher prediction accuracy, robustness and adaptability can be achieved.

Chapter 5.3.1 Design of the user interface

The system provides an interactive dashboard for experts to revise electricity demand predictions produced by a machine learning model. The dashboard displays various information sources, including historical data, model features, and news articles, to support experts in making informed revisions.

Function	Purpose
Visualize predictions	Allows users to view and modify the electricity demand curve predicted by the AI model. This makes it easy for users to see the AI model's predictions and make appropriate adjustments based on them.
Draggable points	Provides an intuitive way for users to directly adjust demand values by dragging points on the curve in the GUI.
Observation operations	Provides users with different types of information to help them make more reasonable decisions. This information includes technique information, modeling information and contextual information.
Global/Local adjustment	Provides two adjustment modes (global and local) according to user needs.
Adjustment confirmation	Allows users to confirm, undo or reset their adjustments before finalizing them.
Result saving	Saves the user's final adjustments and behavior logs for downstream analysis.
Survey	Requires users to fill out questions regarding their adjustment motivations and satisfaction levels. This helps capture users' subjective feelings and motives for their adjustments.

Table 5.2 Function of the user interface

There are three main stages in this human-in-the-loop revision process, and the main functions of this process is shown in table 5.2:

- Preparation: The expert first examines contextual information and technique/model details to understand the context and basis of the predictions. Such background knowledge helps inform subsequent revisions. The expert can inspect historical data, compare between days, and view model features to gain a holistic picture.

- Adjustment: Armed with relevant knowledge, the expert then chooses to either make global or local adjustments depending on needs. For global revisions, the expert specifies a percentage change and time window; for local changes, selectable points allow focused modifications. The expert applies revisions by dragging points or entering a change value.

- Reflection: After applying adjustments, the expert can view feedback comparing the original and revised curves to reflect on the influence of their judgments. When satisfied, the expert confirms changes and fills a survey regarding adjustment motivations and satisfaction.

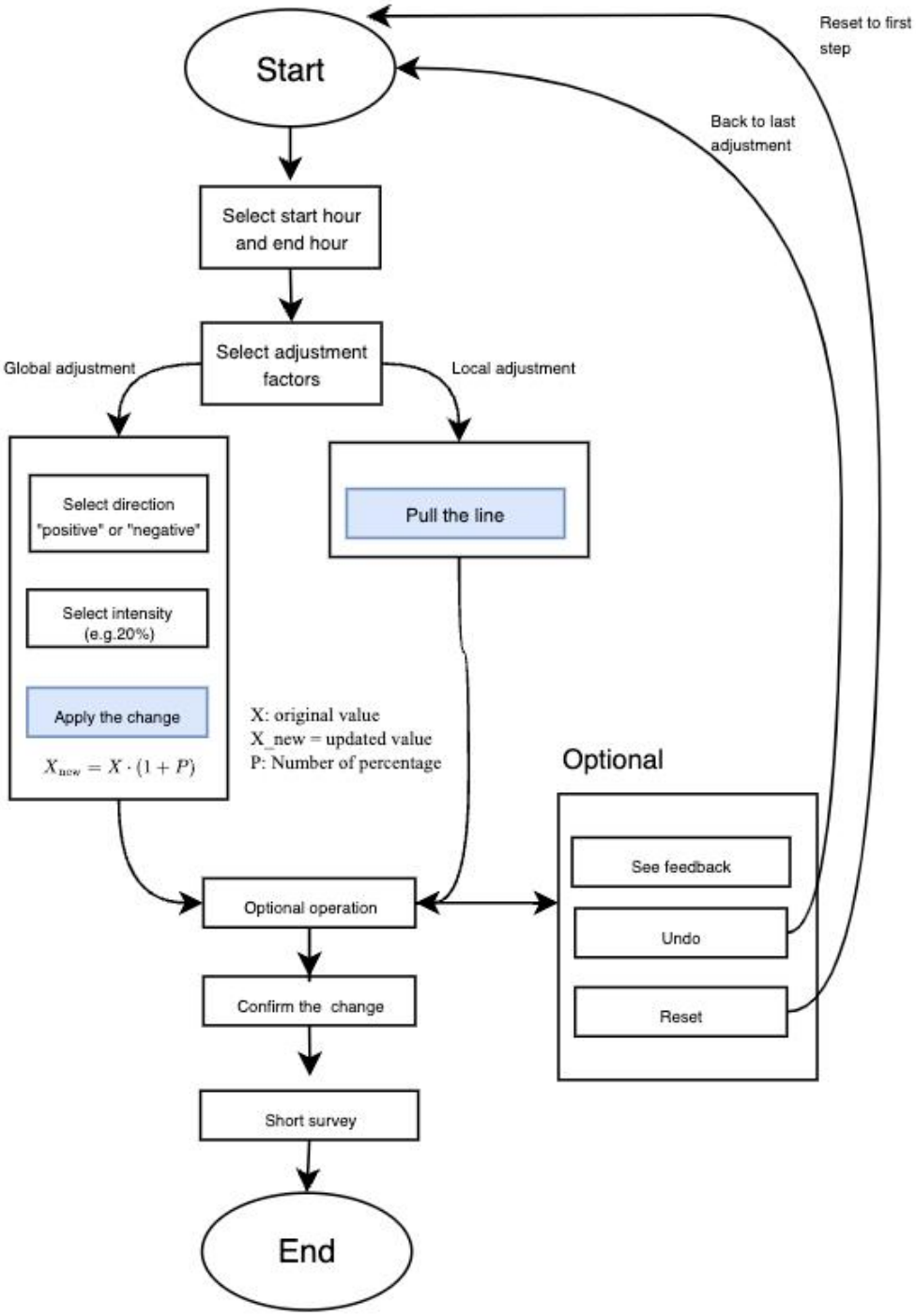

Figure 5.1 The operation flow of the User Interface

The operational process is shown in figure 5.1. At the beginning, the predicted demand curve is visualized, allowing experts to inspect the machine-generated outputs and identify potential improvements. Experts can choose to make either global revisions to the entire curve or local adjustments to select portions, accommodating different human preferences. Experts directly adjust the curve by dragging points to specify new demand values. This intuitive design lowers the barrier for experts to modify the predictions.

After revising, experts confirm their changes and provide feedback via a survey. Confirming adjustments before saving encourages experts to make deliberate changes. The survey offers insights into experts' reasoning behind the revisions. Saved adjustments and behavior logs capture knowledge from experts to improve the machine learning model.

The interactive dashboard facilitates a collaborative human-AI process where experts first prepare by accessing relevant background information. Then, experts revise predictions based on their judgment and areas identified for improvement. Finally, experts reflect on the impact of their changes before the system saves the final adjusted predictions. This "human-in-the-loop" approach combines the strengths of human expertise and AI to produce more robust predictions. The system's visualization, revision options, and feedback mechanisms support a user-centered process where experts can confidently apply their knowledge. Overall, the dashboard provides an effective environment for human-machine collaboration on demand forecasting.

Chapter 6 - Experiment deign

Chapter 6.1 Overview

This chapter details an experiment to evaluate a human-AI collaboration model for electric power demand forecasting. The experiment, bearing the review number 23-80, has been approved by the appropriate ethics review committee. The aims of the experiment are twofold: 1) to assess if the model improves prediction accuracy through human adjustments of AI predictions. 2) to gain insights into optimizing the human-AI collaboration process.

Participants receive training to ensure unbiased results. The experiment consists of preparation, prediction adjustments, and evaluation stages. During adjustments, participants use AI outputs and feedback to optimize predictions while their behaviors are recorded. Data collection and analysis procedures are designed to evaluate the effectiveness of AI assistance and the human-AI collaboration. Analysis of prediction accuracy, adjustment behaviors, and participant feedback are conducted using quantitative and qualitative methods.

It is hypothesized that human adjustments will improve prediction accuracy and that participants will be more confident in their adjusted predictions. The experiment aims to validate these hypotheses and optimize human-AI collaborative systems.

In summary, a comprehensive experiment design is proposed to systematically evaluate the human-AI collaboration model through both accuracy metrics and insights into the collaborative process. The evaluation aims to identify ways to enhance the integration of humans and AI.

Chapter 6.2 Objective of the experiment

The primary goal of our experiment is to assess the effectiveness of our human-AI collaboration model in the context of electric power demand forecasting. Specifically, we aim to validate whether this model, when compared to AI-alone predictions, can enhance the forecasting accuracy through manual adjustments by human users.

Our experiment is not merely designed to evaluate the improvement in prediction accuracy, but it also seeks to provide a deeper understanding of the inherent patterns and dynamics of human-AI collaboration. This study is rooted in the belief that human and AI, when working together, can capitalize on the strengths of each other, thereby surpassing the predictive capabilities of either acting alone. We anticipate that through this experiment, we can shed light on the underlying operation flows of human users when interacting with AI, and understand their decision-making processes.

We aim to answer several questions through our experiment: Can human users generally improve AI predictions in a forecasting task? What is the operational flow during their interaction with AI, and can we trace this back to their decision-making processes? What kind of information do users rely on when making adjustments to AI predictions? And, do factors such as user satisfaction, confidence in the system, and trust in the results influence their adjustments and the overall prediction accuracy?

The insights gained from our experiment would not only validate our human-AI collaboration model but would also contribute to the broader understanding of how best to design and implement such systems in other prediction-based applications.

It is through these objectives that our study endeavors to enhance the integration of AI tools in practical decision-making processes.

Chapter 6.3 Participant group and training

Due to current venue and time constraints, the experiment participants were recruited from the University of Tokyo, with no restrictions on personal background, academic field, or degree. It cannot be assumed that users with a certain domain knowledge would better understand the AI's predictions and make well-grounded adjustments. In reality, not all people have a statistical or domain-specific predictive background. Therefore, the participant group should have varying degrees of familiarity and expertise.

Participants need to perform a series of activities in the human-AI collaborative system, including comprehensive training on system operation—from interpreting historical data and understanding AI predictions to making their own predictive adjustments. This step-by-step training aims to ensure that participants can make full use of the system functions.

It is noteworthy that the current and future participant groups are not limited to industry professionals or domain experts. The goal is to include users with a wide range of backgrounds and expertise levels to increase the diversity of feedback and insights collected in the experiment. This diverse user base will help understand the variability in interactions and decision-making processes among users with different levels of expertise and familiarity with AI tools.

The way participants use the system, their adjustments and feedback will be recorded in detail for analysis after the experiment ends. This dataset will provide valuable information on user-AI system interaction patterns and insights into the user decision-making process. The goal is to explore the effectiveness of human-

AI collaboration in the real world and the potential to expand and improve such collaboration.

Without restricting the participant group to those with professional expertise, the experiment aims to study human-AI collaboration across a diverse range of backgrounds. Comprehensive system training ensures that people of all expertise levels can participate meaningfully. A diverse participant group provides varying perspectives to gain a more holistic understanding of how users interact with and think through AI tools. Broad participation and detailed records of user behavior will enable analyzing how to strengthen human-AI collaboration in reality.

Chapter 6.4 Experiment design and procedure

In the experimental design, three critical stages as shown in figure 6.1 will be conducted to comprehensively evaluate the extent to which AI assists in power market forecasting.

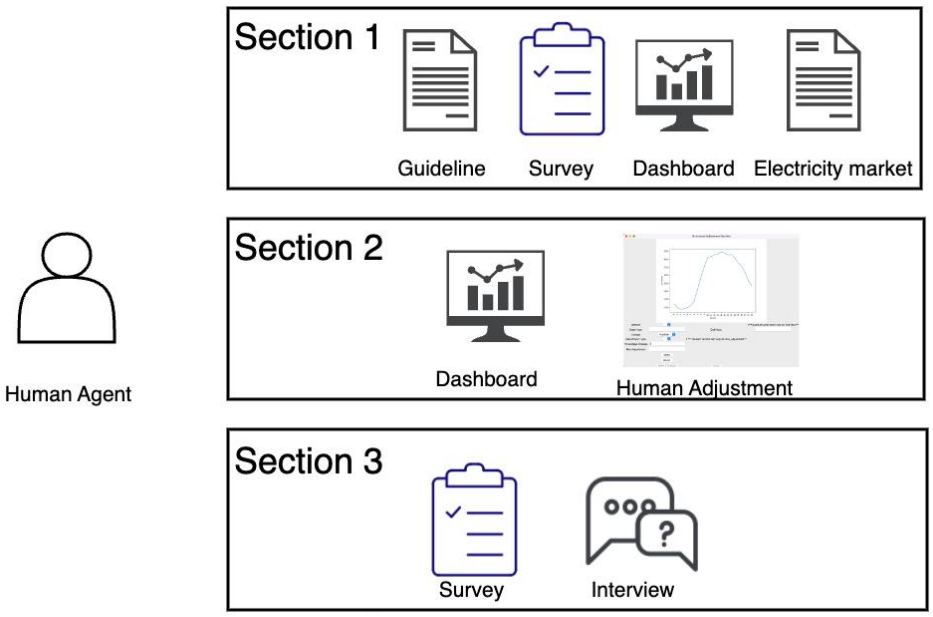

Figure 6.1 The experiment process

The key goal of the first stage is to ensure that participants have sufficient background knowledge to minimize the influence of biases. It will be ensured that participants have adequate background knowledge to participate fairly and comprehensively in the experiment. The purpose of this step is to minimize biases resulting from insufficient prior knowledge and thus reduce their impact on the final results. As part of this process, participants will receive four key documents: 1) An experimental guide describing the entire experimental process and explaining how to effectively use the "interactive dashboard"; 2) A pre-survey to determine demographic information and measure participants' basic understanding of AI-related fields and power markets, thus recognizing the diversity and proficiency of participants. This will include age, education level, and knowledge of AI and power markets; 3) An interactive dashboard as a key tool will provide a dynamic platform to closely examine historical data and insights from generalized additive models (GAMs). It will have various functions such as examining an

overview of historical weather and power data, comparing any two days, and examining power data hourly; 4) "An Introduction to Power Markets" will promote understanding of the factors influencing power demand and emphasize the importance of weather and human activities, including temperature, precipitation, humidity and atmospheric pressure, as well as the concept of changes in power consumption over time, day of week and month. These detailed resources will ensure that participants have a comprehensive and in-depth understanding of the task and background before the start of the experiment.

The key goal of the second stage is to observe how participants use AI to adjust and optimize predictions. Next, the predictive capabilities of AI will be utilized to allow participants to adjust the predictions. AI can identify complex patterns in historical data and generate initial predictions. Users can optimize predictions using the detailed explanatory information and interactive feedback provided by AI. Each adjustment made by the user will be recorded to assess the role of AI in assisting users in adjusting and optimizing predictions. The adjustment process will be carried out through a customized Python user interface with multiple functions such as selecting adjustment factors, determining effective time periods, selecting adjustment methods (adjusting by percentage or maximum value), determining adjustment intensity, comparing results before and after adjustment, selecting undo or confirm changes, and completing multi-choice surveys after each adjustment to capture decision criteria.

In the adjustments, arbitrary adjustments will not be allowed, but a structured method will be adopted. The framework is built on Marmier et al.'s study (Marmier & Cheikhrouhou, 2010). However, this study modifies the framework to provide an easy-to-understand user interface. Information overload and unfriendly user interfaces that could affect accuracy are hoped to be avoided. Importantly, unlike

Marmier et al.'s study, this method allows participants to receive feedback after each adjustment, which other studies have shown can reduce overconfidence and overprediction bias, thereby improving accuracy. After each adjustment, participants will need to complete the multi-choice survey shown in the figure to capture decision criteria at each stage, providing data for subsequent analysis.

The key goal of the third stage is to comprehensively evaluate the effectiveness of AI in assisting with predictive adjustment and optimization. Finally, detailed feedback from participants will be collected to evaluate the effectiveness and extent of AI assistance. Views of participants on the role of AI in assisting with predictive adjustment and optimization will be surveyed. If participants believe that the detailed explanatory information and interactive feedback provided by AI are critical to optimizing and improving their predictions, this indicates that AI has played a key role in this process. Informal interviews will also be conducted to further understand the reasons behind the participants' decisions and the degree of adjustment. This step aims to gain insight into human-AI collaboration from participants.

AI assists in optimizing predictions by providing deep pattern identification of historical data and detailed explanations. It can generate initial predictions and then continuously optimize and improve predictions through interactive feedback to make predictions more accurate. Therefore, it is expected that the explanatory information and interactive feedback from AI will be critical for participants to adjust the final prediction results. If participants' feedback confirms this, it indicates that AI has successfully played a key role in the prediction process, facilitating efficient collaboration between humans and AI.

Through this experimental design, an in-depth exploration of the extent to which AI assists in power market forecasting is hoped for. It will be evaluated whether

participants' predictive adjustments significantly improve the accuracy of predictions, and the effect of AI's detailed explanatory information in this process. In addition, a comprehensive understanding of participants' perceptions of the extent to which AI assists in this process will be gained. Through this experiment, an in-depth understanding of the human-AI collaborative model is expected to optimize this process and advance the application of AI in power market forecasting.

Chapter 6.5 Experiment data collection and analysis procedure

First, various data will be collected in the data collection phase. In this subsections, characteristics of different types of data is introduced. Then the data processing procedures and data analysis methods to deal with those experiment data are further introduced.

Dataset	Data Information	Data Characteristics	Data Structure
Behaviour Log	Records the timestamp and action performed by a user interacting with a system	Two columns: time (string) and behaviours (string)	Dataframe with two columns: time and behaviours
Each Adjustment	Contains numerical data with multiple columns and rows, each row represents a specific timestamp	Multiple columns of float numbers	Dataframe with multiple columns containing float numbers
Final Adjustment	Contains pairs of x and y values representing adjustments to a model or a graph	Two columns: x (float) and y (float)	Dataframe with two columns: x and y
Survey	Contains survey results (information sources, satisfaction level, reason, better than AI, improvement)	Five columns with various data types (list, int, string)	Dataframe with five columns

Table 6.1 The data information of the collected from experiment

The datasets as shown in table 6.1 contain information to analyze participants' adjustments to AI predictions for electric power demand, and there four different dataset were collected after each experiment. The behavior log records each participant's actions while using the system, giving insight into which information they found useful and which features they used the most. The adjustment CSVs contain the initial AI predictions and the participants' adjustments to those predictions at different points in time. Comparing the initial and adjusted values can show how participants changed the predictions and by how much. The final adjustment CSV contains only the final adjusted predictions, showing the end results after participants reviewed the information and made their changes. The survey CSV contains information about the sources participants found most helpful, their satisfaction levels, reasons for adjustments, and whether they felt their adjustments improved upon the AI predictions. Analyzing the data structures, with their timestamps, numerical prediction values, and text responses, can reveal how participants collaborated with the AI system to potentially improve the final predictions. Overall, the datasets provide a holistic view of how end users interacted with the system and adjusted the AI predictions, which can help evaluate the effectiveness of the human-AI collaborative model.

In the data processing stage, preprocessing of the collected data will first be needed, including handling missing values, cleaning invalid data, encoding text data, etc. Take the participants' prediction results as an example. The same preprocessing operations will be performed on their prediction results at different points in time.

In the data analysis stage, the differences between the participants' adjusted predictions and the AI's initial predictions will be mainly considered. Secondly, to

gain a deeper understanding of the human-AI collaborative process, the behaviors taken by participants in predictive adjustments, such as adjustment frequency, adjustment magnitude, and adjustment methods used, will be analyzed. This will help understand how participants used the information provided by AI to adjust predictions.

In addition, feedback provided by participants during the experiment, including evaluations of AI, satisfaction with the experiment, and confidence in predictions will be analyzed. Qualitative text will be encoded and quantified to analyze this feedback. This will help understand the actual effects of AI in assisting power market forecasting. This feedback comes from brief surveys after each change confirmation by participants and questionnaires after the entire experiment.

Considering the limited sample size, if possible, statistical methods will be used to analyze these data, such as descriptive statistical analysis, correlation analysis, and regression analysis. Descriptive statistical analysis will be used to analyze participants' qualitative feedback, and correlation analysis and regression analysis will be used to analyze participants' predictive adjustment behaviors and results.

The goal of this stage is to understand the effectiveness of AI in assisting power market forecasting and the human-AI collaborative model through in-depth data analysis. These two goals are closely related, with the former focusing on the effects of AI and the latter focusing on the human-AI collaborative process. They are equally important to the research.

Chapter 6.6 Expected results

In designing this study, the goal is to verify that by introducing the human factor when AI cannot make good predictions due to lack of relevant information, better results can be achieved. Secondly, a deeper understanding of the collaborative model between AI and humans in electric power demand forecasting and how this collaboration affects the accuracy of prediction results is also hoped for.

The following are expected to be observed:

- Participants' predictive adjustment behaviors: By analyzing participants' behavior logs in the system, some common adjustment patterns and trends are expected to be found. For example, certain adjustment methods may be used more frequently, or in certain situations, participants may tend to make large-scale adjustments.

- The extent to which AI assists in the prediction process: It is expected that by comparing the accuracy of AI predictions and participants' adjusted predictions, the contribution of AI in the prediction process can be understood. If participants' adjustments can significantly improve the accuracy of predictions, this will prove the effectiveness and degree of AI assistance.

- Participants' feedback and satisfaction with AI: It is expected that through participants' feedback surveys, their evaluations of AI, satisfaction levels, and confidence in predictions can be understood. This will help understand the acceptability and feasibility of AI in actual applications.

- ❖ The human-AI collaborative model: Finally, by analyzing participants' adjustment behaviors and feedback, the human-AI collaborative model in electric power demand forecasting is hoped to be revealed. This will help understand how to more effectively design and optimize AI systems to improve their performance and efficiency in power market forecasting.

The following are some expected results, which are also the hypotheses of the research:

- ✓ ***H1:*** *It is hypothesized that users' adjustment behaviors have a significant impact on AI predictions. By integrating the human factor, the accuracy of predictions can be significantly improved compared with purely machine-generated results.*

- ✓ ***H2:*** *Compared with AI-generated responses, the adjustment results from participants with technique background will be better than the adjustment results form participants without technique background.*

In the hypothesis statement, the "adjustment results" refer to the results of human adjustments made to the original machine forecasting results. Additionally, the "technical background" refers to whether the participant has an AI-related background, such as being a computer science major or having work experience in the field. This research is interested in investigating whether there is a difference in the performance of participants with "technical background" and those without. All of these expected results will be tested after the experiment using the collected data.

This study is hoped to provide in-depth insights to further understand the role of AI in this process and evaluate AI's contribution to improving the accuracy of predictions.

Chapter 7 - Results and Discussion

Chapter 7.1 Overview

This chapter examines whether human adjustments with different levels of technical background can improve machine learning predictions of electricity demand. Preliminary results show that when machine predictions falter, human adjustments - particularly those with technical expertise - can provide benefits. Analyses of participants' interaction logs reveal that those with technical backgrounds actively seek information before making incremental changes, while non-technical participants quickly make adjustments to finish the task. The findings suggest that technical knowledge may allow humans to better complement AI, though the small sample size limits generalization. The developed framework for visualizing interaction processes provides insights into human roles in human-AI collaboration, indicating technical expertise as an important factor for improving AI forecasting outcomes through human adjustment. More research with larger samples is needed to validate these findings and illustrate how technical expertise influences human adjustment processes and outcomes.

Chapter 7 begins by outlining the objectives and information contained in the collected datasets from the experiment (Section 7.2). Four types of data are described: behavior logs, CSV files per adjustment, final adjustment CSV, and survey CSV. Preprocessing steps taken to ensure data quality are also summarized. Chapter 7.3 then provides an initial comparison of human-adjusted results versus pure machine predictions. Plot visualizations demonstrate how human adjustments vary in aligning with true demand data versus the original statistical forecasts. Chapter 7.4 presents the main results and discussion. Participants are classified into technical and non-technical groups based on background. Statistical tests examine hypotheses that human adjustments improve on machine predictions, especially during poor forecasting periods, and that technical users outperform non-technical

ones. Plots visualize the error of adjustments made. Chapter 7.5 delves into analyzing user interaction logs to uncover process differences between technical and non-technical groups. Visualizations and analysis reveal how those with technical expertise actively seek information and incrementally refine adjustments. Finally, Chapter 7.6 proposes a visualization framework to trace user behaviors during human-AI collaboration. Transition plots with order and probability indicators showcase how this framework can reveal user preferences and decision-making processes based on their interactions.

Chapter 7.2 Objective and information of collected data

In our study, we collected four types of datasets that provide detailed information on how participants adapted AI to predict electricity demand. Below, we describe the information and characteristics of these datasets specifically, as well as our data processing methods.

Dataset	Data Information	Data Characteristics	Data Structure
Behaviour Log	Records the timestamp and action performed by a user interacting with a system	Two columns: time (string) and behaviours (string)	Dataframe with two columns: time and behaviours
Each Adjustment	Contains numerical data with multiple columns and rows, each row represents a specific timestamp	Multiple columns of float numbers	Dataframe with multiple columns containing float numbers
Final Adjustment	Contains pairs of x and y values representing adjustments to a model or a graph	Two columns: x (float) and y (float)	Dataframe with two columns: x and y
Survey	Contains survey results (information sources, satisfaction level, reason, better than AI, improvement)	Five columns with various data types (list, int, string)	Dataframe with five columns

Table 7.1 Information of collected data from the experiment

- Behavior Logs: This dataset records behavior logs of user interactions with the system, including timestamps and actions performed by the user. The data is in CSV format and contains two columns: time (string, representing the timestamp) and behavior (string, representing the action performed). The data structure can be represented as a data frame containing both time and behavior columns.
- CSV per adjustment: This data set contains numeric data with multiple columns and rows. Each row represents a specific timestamp and each column contains the value of that timestamp. The data is in CSV format and contains

multiple columns of floating point numbers. The data structure can be represented as a data frame containing multiple columns of floating point numbers.

- Final Adjustment CSV: This data set contains pairs of x and y values that represent model or graph adjustments. The data is in CSV format and contains two columns: x (floating point) and y (floating point). The data structure can be represented as a data frame containing two columns, x and y.
- Survey CSV: This data set contains the survey results, including the source of information, satisfaction, reason, whether it is better than AI, and the percentage of improvement. The data is in CSV format and contains five columns: source (list of strings), satisfaction (integer), reason (string), whether better than AI (string, "Yes" or "No"), and improvement percentage (integer). The data structure can be represented as a data frame with five columns containing the information source, satisfaction, reason, whether it is better than AI, and improvement percentage.

In the data processing phase, we performed the following steps to ensure the quality of the data and the accuracy of the analysis: (1) Data cleaning: We first removed all invalid data points, such as those generated due to user error operations or system errors. (2) Data format conversion: The data collected in the experiments need to be converted to some extent before being used for analysis. For example, the short survey in the User interface is used to collect data of int or float data type. (3)Mutual integration of data sets: The data sets are incorporated in a targeted manner. For each operation logic of the user or whether it is convenient to record data, we only divided the data set into four. But sometimes with data analysis, the information given by a single dataset is limited, such as the

behavior_log.csv set each_adjustment_csv before we can know what kind of user patterns the user experienced before going through the current adjustment.

During the data analysis phase, we took the following approaches to understand the data in depth and to draw conclusions. (1) Descriptive statistical analysis: We first performed descriptive statistical analysis, including calculating the mean, median, and standard deviation, to obtain the basic information and distribution characteristics of the data. (2) Variable relationship analysis: We analyzed the relationship between each variable, including correlation analysis, covariance analysis, etc. Through these analyses, we can understand which variables have a greater impact on people adjusting their forecasts. (3)Hypothesis testing: We set some research hypotheses and tested them by t-test, ANOVA and other statistical methods.

Chapter 7.3 Comparison of human adjusted results and pure machine prediction

This study aims to address the limitations of machine predictions due to real-world factors that prevent machines from acquiring or learning from data relevant to accurate predictions, especially for rare events. If faced with rare events, machine learning prediction algorithms may struggle to make accurate forecasts. Considering humans can capture, integrate and infer from latent information, we consider integrating human factors to complement machine learning's informational limitations to improve accuracy while enhancing the robustness and resilience of machine and statistical learning algorithms in special situations. Though the study's sample consists of only nine participants so far, insufficient to represent the population, it still demonstrates some interesting information.

The nine participants had diverse backgrounds. The data from the pre-survey shows that the participants are relatively young, with an average age of 26.44 years. They have a wide range of educational backgrounds, with 4 participants having a Master's degree, 3 having a Ph.D. degree or higher, and 2 having an undergraduate degree. The participants' majors are also diverse, with 2 majoring in System Architecture, 2 in Computer Science/Data Science/Statistics/Mathematics or AI-related major, and the others each having a different major. All 9 participants reported that they do not have a basic understanding of the electricity market of Japan.

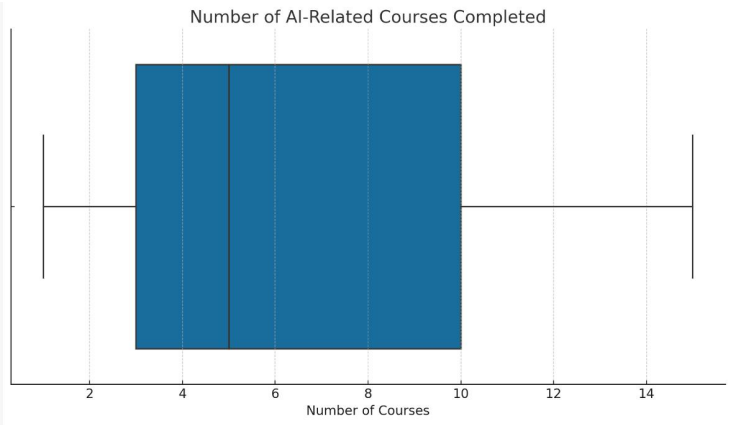

Figure 7.1 Number of AI-related courses completed

This box plot in figure 7.1 shows the distribution of the number of AI-related courses completed by the participants. The median (the line inside the box) is around 5 courses. The box represents the interquartile range (from the 1st quartile to the 3rd quartile), and the whiskers represent the range of the data. There are no outliers in this data.

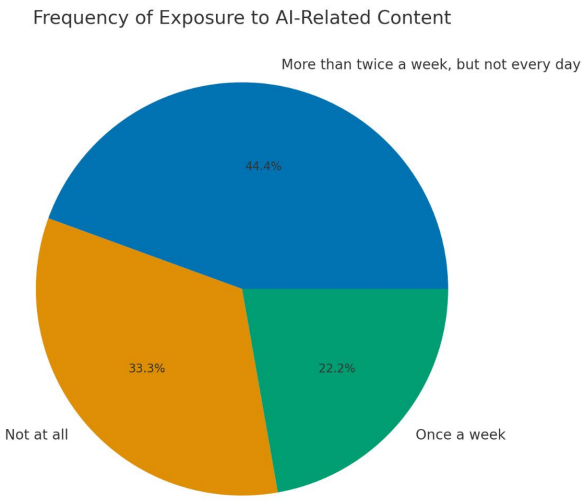

Figure 7.2 Frequency of exposure to AI-related content

As shown in figure 7.2, the participants are exposed to AI-related content at varying frequencies, with 4 participants being exposed more than twice a week but not every day, 3 participants not being exposed at all, and 2 participants being exposed once a week.

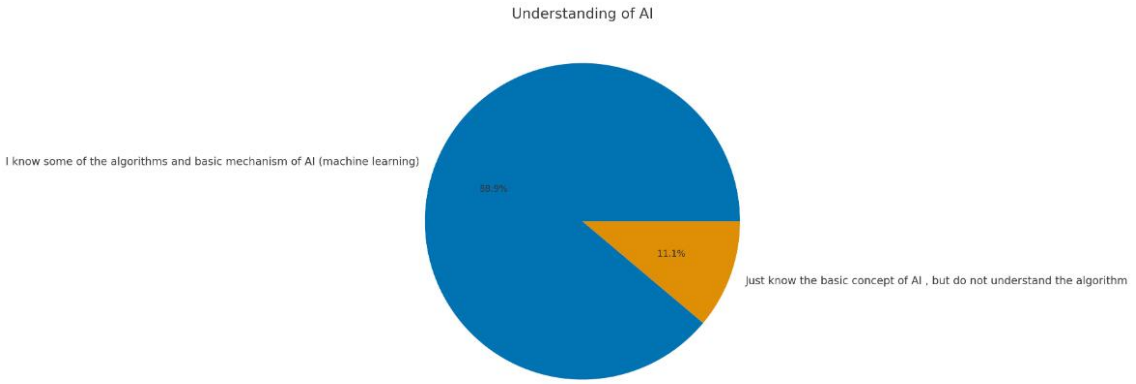

Figure 7.3 The proportion of understanding of AI

In terms of their understanding of AI as shown in figure 7.3, 8 participants report that they know some of the algorithms and basic mechanisms of AI

(machine learning), while 1 participant just knows the basic concept of AI but does not understand the algorithm.

Taking into account the research goal, the participants were classified into two groups: Group 1 with AI background and Group 2 without AI background, based on the number of AI-related courses they had completed in university or online classes. AI-related courses include machine learning, data mining, statistics, and mathematics. After a simple calculation, all participants who had completed five or more AI-related courses were classified as Group 1, and those who had completed fewer than five AI-related courses were classified as Group 2. Ultimately, six participants were classified as Group 1 and three as Group 2.

Based on the classification of participants, the results and implications will be discussed in three sub-sections:

- Sub-section 1: We aim to demonstrate that human intervention can improve machine predictions,a key aspect of our research. This study assumes participants with technical backgrounds handle information better than those without, so we also investigate whether Group 1 makes better human adjustments than Group 2. We examine two hypotheses:

  - ***Hypothesis 1****: In next-day predictions, the accuracy of human adjustments exceeds pure machine predictions.*
  - ***Hypothesis 2****: In next-day predictions, the RMSE of Group 1's final human adjustments is lower than Group 2's.*

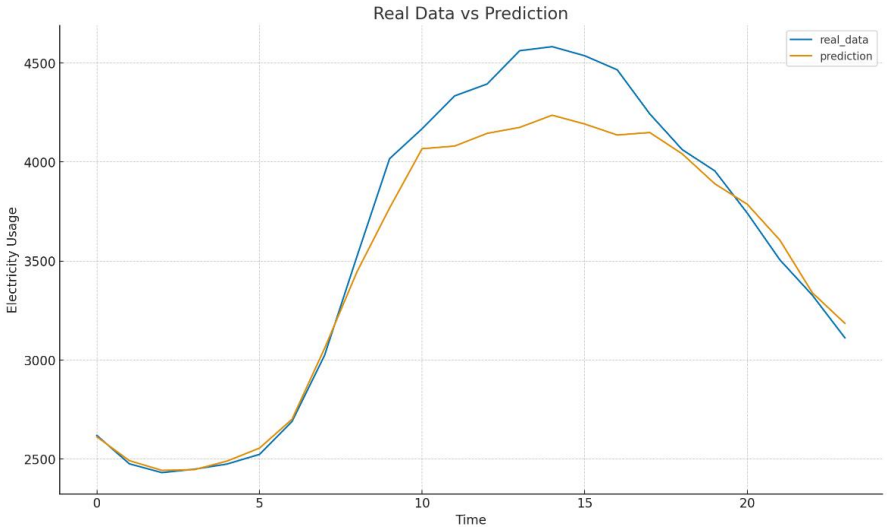

Figure 7.4 Comparison between real data and prediction result

The plot of the difference between real electricity data and machine-predicted electricity data shows that the forecast performs poorly during peak usage. As one of the goals of this research is to investigate whether human adjustments can improve machine-generated forecasts, the research examines whether participants can contribute knowledge or information during 9 am to 4 pm to optimize forecasts. Two hypothesis tests with limited time ranges are tested:

- ✓ ***Hypothesis 3****: During 9 am to 4 pm, the mean difference of the prediction accuracy of human adjustments exceeds pure machine predictions.*
- ✓ ***Hypothesis 4****: During 9 am to 4 pm, the mean difference of Group 1's RMSE of final human adjustments is lower than Group 2's.*

- ❖ sub-section 2: while previous studies have demonstrated that experts' information not contained in models can improve time series forecasting accuracy, the studies only reported results and authors' interpretations without

much explanation of the mechanisms and principles behind why humans can increase model forecasting accuracy. Though one study mentioned technical background was not a significant factor impacting results while providing contextual information was, it still only reported results without mechanistic or principled explanations for why technical background did not affect final predictions.

This study argues attention should not only be on results but also on processes. With human factors incorporated, results are influenced by multiple factors not necessarily causally related to initial assumptions. Though technical background may not significantly impact results, we should examine process differences rather than solely outcomes in human-AI collaboration.

This study aims for results approximating real environments, therefore in reality, tasks are completed by non-experts in most cases, as in our electricity forecasting task where not all participants had technique backgrounds. The research seeks to understand how those with and without technique backgrounds manifest differently, if at all, in performing an electricity prediction task. Overall, this section presents data analysis from two angles:

- Participants with technique background versus those without AI background
- Comparing adjustments improving versus worsening prediction accuracy relative to machine forecasting

In this section, the research will analyze operational processes underlying those adjustments, examining differences for participants with technical backgrounds

and commonalities among adjustments outperforming machine forecasting, potentially indicating patterns.

- sub-section 3: this section builds on sub-section 2's discussion. Fundamentally, this study investigates and tracks participants' performance in problem-solving (improving machine forecasting results for electricity prediction) based on the theoretical assumption that humans complement AI roles in interaction. AI utilizes its advantages in processing vast high-dimensional data and analyzing underlying meanings to provide useful information, helping humans understand contextual and other noteworthy information. Upon information reception and understanding, humans apply their capabilities for abstract information integration, divergent thinking, and inference to provide information missing from machine forecasting. This study provides a quantitative framework atop sub-section 2's discussion that tracks users' procedural workflows, tracing decision-making processes. This framework observes not only how different users operate but also how those with and without technical backgrounds operate, potentially revealing patterns.

Chapter 7.4 Main results and discussion

In this section, we aim to investigate and demonstrate two crucial aspects of our research: 1) whether human adjustment leads to better performance than machine forecasting, and 2) whether the participants in Group 1 (with a technical background) outperform those in Group 2 (without a technical background).

Firstly, based on the measurement before, I classified participants into 2 groups based on their survey answers:

- Group 1(with technique background): User_2, User_3, User_4, User_6, User_7, User_8
- Group 2 (without technique background): User_1, User_5, User_9

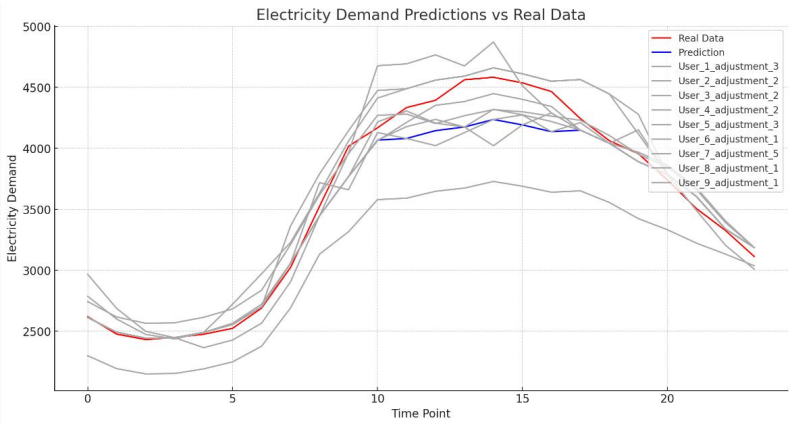

Figure 7.5 Human adjustment results, prediction and real data

The figure 7.5 presents a comparison of electricity demand predictions, the final human adjustments to those predictions, and the actual observed electricity demand. Each line represents a different data series. The red line represents the actual observed electricity demand ("Real Data"). This is the true value that the predictions are trying to approximate. The blue line represents the initial predictions made by the machine learning model ("Prediction"). These are the outputs before any human adjustments were made. The dark grey lines represent the final adjustments made by each of the nine users. Each user made several adjustments over time, and these lines represent their final adjustments.

It shows, excluding the participants' adjustments, the forecasted data is lower than the actual data. Although the forecasting for the early morning to late morning and afternoon to evening time periods remains highly accurate, the model's predictions are not entirely accurate due to the limitations of historical data, even when the model is aware of the weather for the following day. Moreover, some users' final adjustments align more closely with the real data than others. This could suggest that some users were more successful than others in improving the predictions, although it would be important to assess this over many different time points to draw any firm conclusions. However, the final adjustments made by the users show considerable variability. This suggests that different experts have differing views on how to improve the machine learning model's predictions. The reasons for these differences could be interesting to explore further, as they could reflect different expertise, perspectives, or strategies among the users.

Furthermore, as shown in figure 7.6, it can be observed that after incorporating human intervention, we can see that the red baseline represents the machine forecasting results, and most of the bars in the bar charts are below this baseline. This suggests that the majority of participants have achieved improvements over the original machine forecasting results.

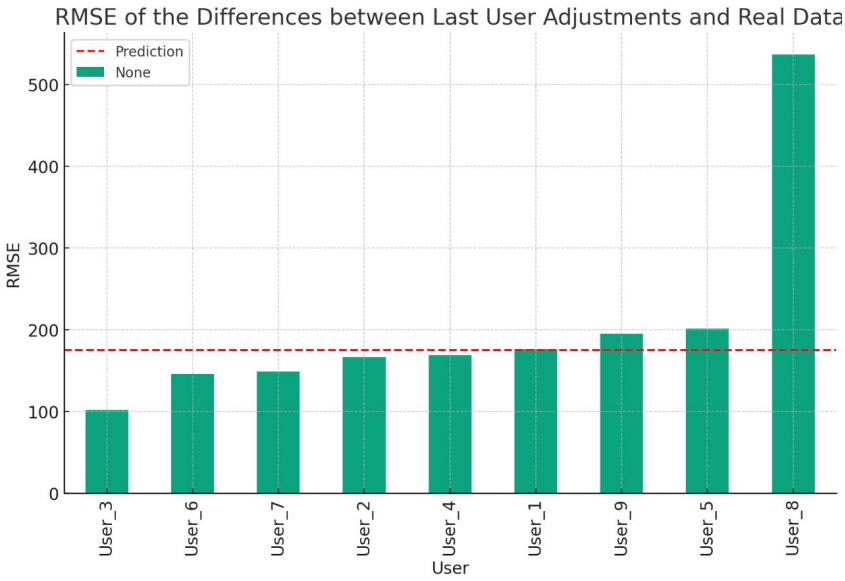

Figure 7.6 RMSE of the differences between human adjustment and real data

Due to the limited size of the dataset, we cannot draw definitive conclusions. However, within our dataset, User 1, User 5, and User 9 have been classified as Group 2, i.e., those without an AI background. Their results show that they worsened the machine forecasting results. In other words, they did not optimize the statistical forecasting, but even made it worse. User 1's result is close to the statistical forecasting, even though he lacks a background in AI-related disciplines, he has experience in physics-based modeling. On the contrary, those participants with technical backgrounds have much better results than statistical forecasting. Noticeably, since User 8 completely misunderstood the contextual information, he adjusted the forecasting results in the opposite way. Therefore, he is an outlier, and I will exclude him from this research.

	User_3	User_6	User_7	User_2	User_4	Prediction	User_1	User_9	User_5	User_8
RMSE	101.53	145.73	148.91	166.20	169.24	175.09	176.12	195.21	201.24	536.85

	User_3	User_6	Prediction	User_2	User_1	User_9	User_4	User_7	User_5	User_8
MAPE	2.01%	2.48%	2.89%	3.09%	3.13%	3.24%	3.51%	3.82%	4.14%	12.98%

Table 7.2 RMSE and MAPE metrics of human adjustments and prediction

The table 7.1 provides a comparison of the RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) of the final human adjustments and original machine predictions against the real data. In terms of RMSE, User_3 has the smallest error, with a value of 101.53, indicating that their adjustments are closest to the real data. This is followed by User_6 and User_7, with RMSEs of 145.73 and 148.91, respectively. The machine predictions have a mid-range RMSE of 175.09, meaning they are generally more accurate than the adjustments of User_1, User_9, User_5, and especially User_8, who has the highest RMSE of 536.85. Looking at MAPE, we again find User_3 performing best with the lowest error of 2.01%, closely followed by User_6 with a MAPE of 2.48%. The machine predictions have a slightly higher MAPE of 2.89%, but still outperform User_2, User_1, User_9, User_4, User_7, User_5, and especially User_8, who has the highest MAPE of 12.98%. It's interesting to note that despite the statistical forecasting result's relative accuracy, there is significant variability in the accuracy of the human adjustments. This underscores the complexity of electricity demand forecasting and the potential value of diverse human expertise in improving prediction models.

Furthermore, from figure 7.7 we can see from the heat map compared to the gap between prediction and real data, we are interested in whether participants have improved or worsened the prediction, specifically at what time of the day from 0 to

24 hours, improved or worsened.The horizontal coordinate represents each hour and the vertical coordinate represents the user. In the heat map, blue color indicates that the difference is positive and the user's adjustment is closer to the actual data than the original prediction. Red color indicates that the difference is negative and the user's adjustment is farther away from the actual data. The shade of the color indicates the size of the difference; the darker the color, the larger the difference.This heat map facilitates visualizing patterns and trends in the data. For example, we can see that most users can improve the AI results when the prediction is poor (9-14), but some users may also overreact at other times, or add their own subjective bias, making the machine forecasting results worse.

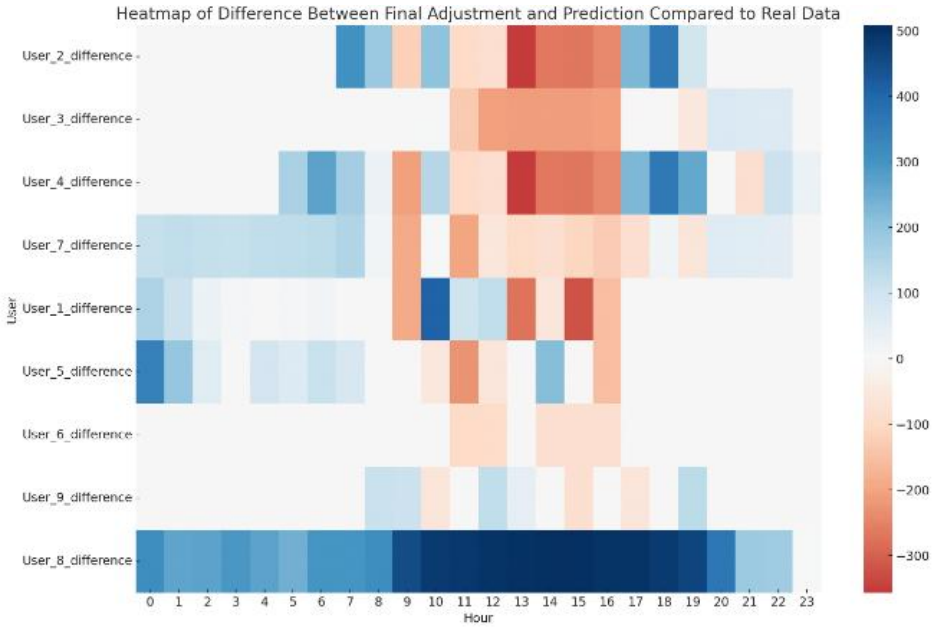

Figure 7.7 Heatmap of the difference in human adjustments and prediction compared to real data

From results and visualizations above, it is clear to know that human can improve the statistical forecasting results. Statistical tests is important to prove our findings.

H1: After excluding user 8, conducting a one sample t-test for hypothesis 1 yielded a t-statistic of -1.08 and a p-value of 0.316, indicating no significant evidence that human adjustments improve machine forecasting overall.

H2: However, limiting the analysis to times when machine forecasting performed poorly (9am-4pm) revealed a t-statistic of -2.75 and p-value of 0.029, suggesting human adjustments can improve accuracy when machine forecasting falters.

The study failed to demonstrate a statistically significant overall difference between statistically generated forecasts and human-adjusted forecasts. However, under the constraint of restricting the analysis to time periods (9am to 4pm) when machine predictions were poor, human adjustments were found to provide some benefit, regardless of the individuals' technical background. Due to the small sample size, the results lack generalizability, but they reaffirm that human input is meaningful when machine predictions falter. However, human bias may be harder to control and predict than expected, and human adjustments tend to overcompensate for good machine predictions, thereby yielding worse outcomes.

Although earlier research found that technical background had no meaningful impact on the accuracy of forecast adjustments (Sanders & Ritzman, 1992), this study posits that: 1) those studies are at least 10-20 years old and may lack external validity today, 2) prior experiments were conducted in laboratory settings devoid

of an authentic information processing and feedback cycle as participants were directly asked to make predictions, and 3) advances in AI knowledge, algorithms and visualizations over the past decades could differentially influence how those with and without technical backgrounds approach prediction tasks, limiting the applicability of prior findings. Thus, the current study thus aims to examine whether technical background influences the accuracy of human adjustments to machine-generated forecasts.

H3: The independent samples t-test was conducted to compare the mean RMSE of Group 1 and Group 2. The test statistic was -2.61, which is a negative number. This indicates that the mean RMSE of Group 1 is lower than the mean RMSE of Group 2. The p-value was 0.040, which is less than 0.05. The result of the independent samples t-test shows that the mean RMSE of Group 1 is statistically significantly lower than the mean RMSE of Group 2. This means that we have sufficient evidence to show that the final adjustments of Group 1 users are closer to the actual data than the final adjustments of Group 2 users.

H4: The independent samples t-test was also conducted for the 9am to 4pm time period. The test statistic was -4.55, which is a negative number. This indicates that the mean RMSE of Group 1 is lower than the mean RMSE of Group 2. The p-value was 0.0039, which is far less than 0.05. This indicates that we can reject the null hypothesis, which is that there is no significant difference between the RMSE of Group 1 and Group 2. The result of the independent samples t-test for the 9am to 4pm time period shows that within this time period, the mean RMSE of Group 1 is statistically significantly lower than the mean RMSE of Group 2. This means that we have sufficient evidence to show that within this time period, the final adjustments of Group 1 users are closer to the actual data than the final

adjustments of Group 2 users. This conclusion is consistent with our overall observations.

This result shows that within this time period, the mean RMSE of Group 1 is statistically significantly lower than the mean RMSE of Group 2. That is, we have sufficient evidence to show that within this time period, the final adjustments of Group 1 users are closer to the actual data than the final adjustments of Group 2 users. This conclusion is consistent with our overall observations.

In these two statistical tests, significant results were obtained. These results demonstrate that technical background is a relevant factor affecting people's final judgments, at least in the current research. Although there are sufficient reasons to explain why the results differ from those of previous studies, these are based on personal interpretations. It is hoped to gain insights into whether people with and without technical backgrounds exhibit different behaviors when interacting with AI, through analyzing some operational processes during human-AI interactions. Therefore, in the next part of the research, an in-depth investigation into the interaction process is intended by collecting questionnaire data combined with the data of users' adjustments during interactions.

Chapter 7.6 Data analysis results of Human-AI collaboration

In this study, humans are required to interact with AI to accomplish a task, specifically, to better predict electricity demand, which is a forecasting task. Participants can reference information from data analysis (historical data), modeling information (GAM model), contextual information (weather and news), and their own judgment. They can then make adjustments to the original prediction

using global and local change methods. At this stage, we provided users with ample freedom to operate without restricting their actions. We aim to examine the differences or insightful ideas in the interaction patterns between users and AI in forecasting tasks, laying a foundation for our future human-AI interaction experiments.

This research wants to investigate two scenarios:

- Participants with technical and non-technical backgrounds
- Adjustments with improvements and adjustments without improvements

First, we will conduct some descriptive statistical analyses, we begin by examining the differences between individuals with and without technical backgrounds.

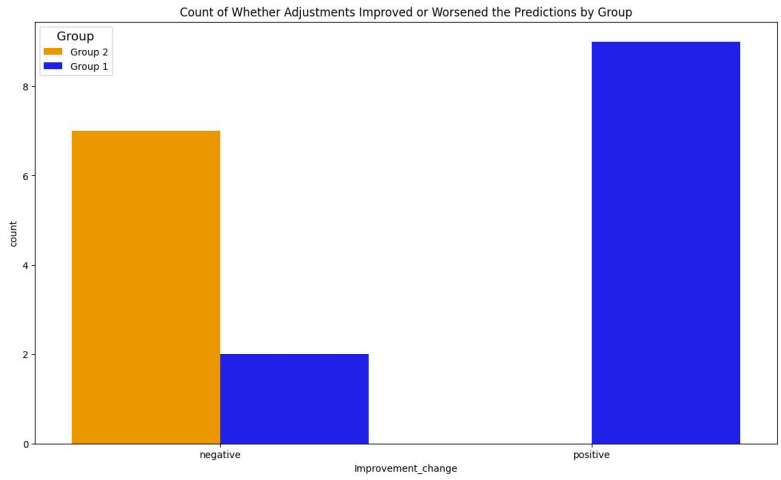

Figure 7.8 Count of whether adjustments improved or worsened the predictions by group

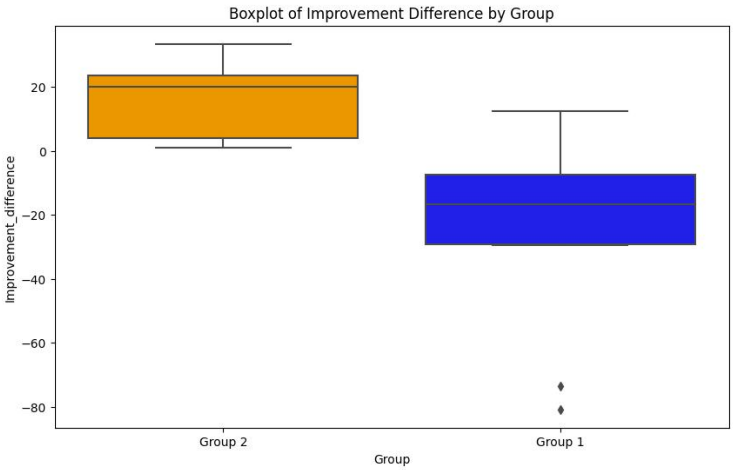

Figure 7.9 Boxplot of difference of improvement in adjustments by group

Based on the figure 7.8 and 7.9, people with a technical background were able to make adjustments that were better than AI almost every time, but people without a technical background made adjustments that made the predictions worse almost every time. The results also show that participants in Group 1 were able to reduce RMSE, while Group 2 increased the error instead.

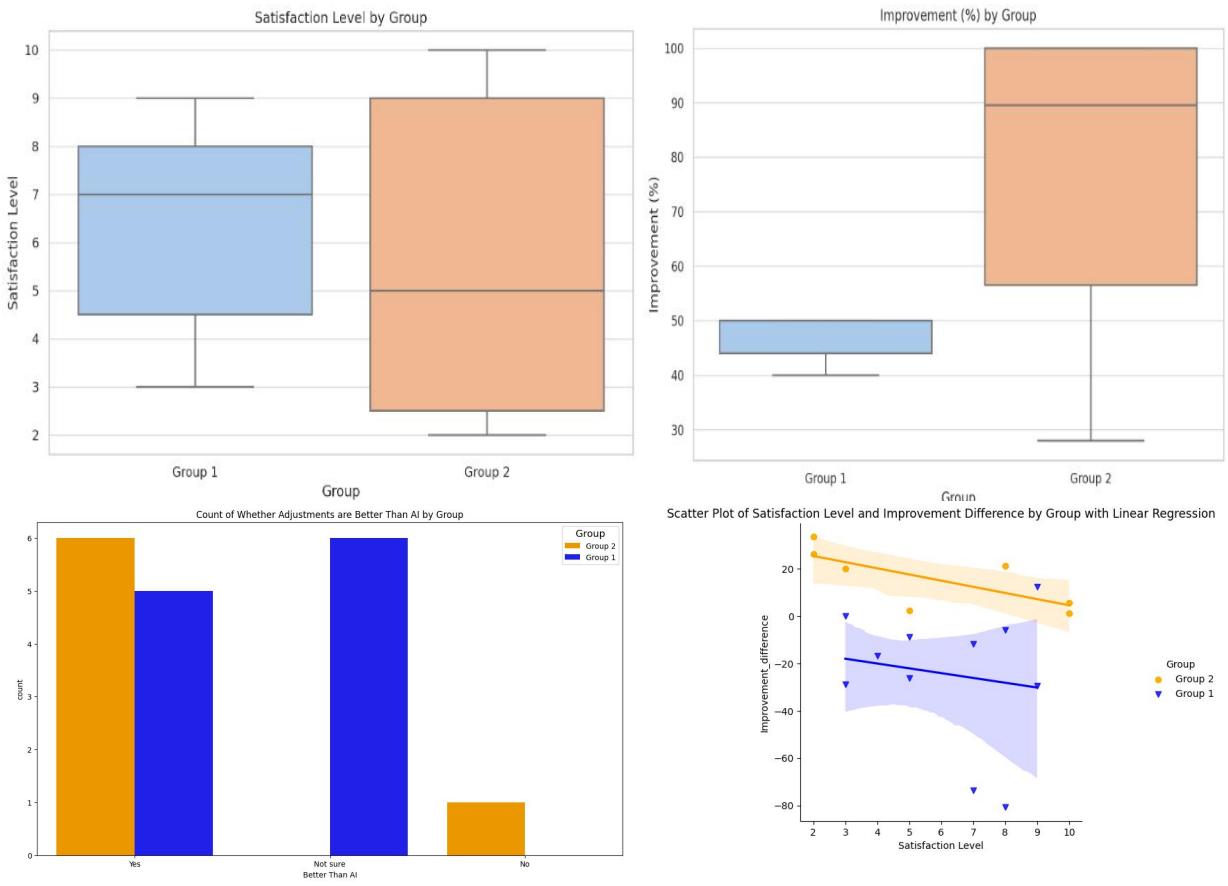

Figure 7.10 Overview of the relationship of satisfaction and improvement in adjustments.

Based on their satisfaction and confidence levels. From the figure 7.10, we can see that the average satisfaction of Group 1 (AI background) users is slightly higher than that of Group 2 (non-AI background) users. This may indicate that AI background users are more satisfied when using the system to adjust power demand forecasts. Non-technical background users are more confident in the results of their adjustments and more easily feel that their adjustments will exceed the machine forecasting results, but technically background users are more careful and cautious, and are more unsure about whether their adjustments are better than the machine forecasting results. Worth mentioning is that the improved satisfaction

seems to be negatively correlated with RMSE, which means that to a certain extent, users can correctly perceive whether they can make better or worse changes.

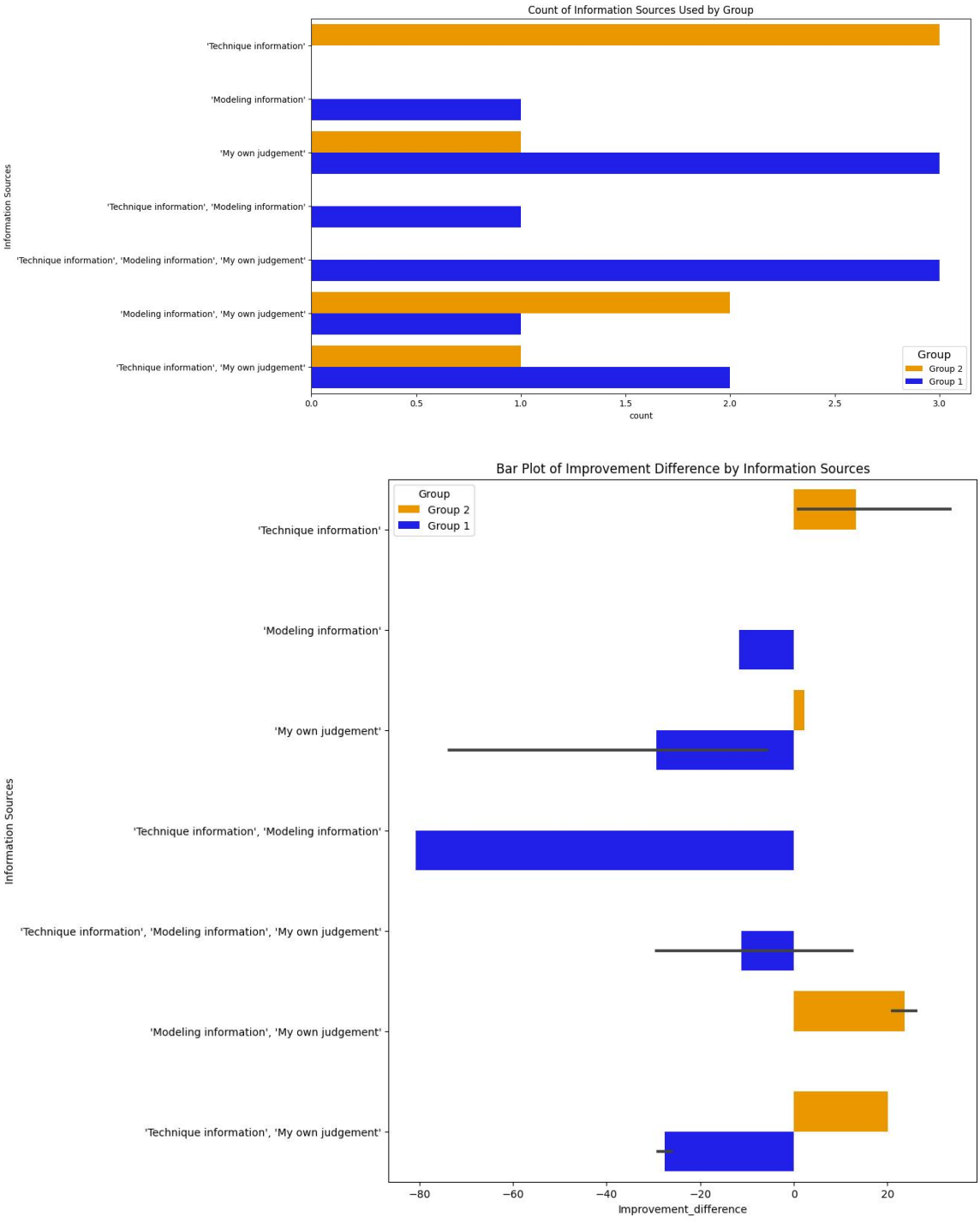

Figure 7.11 Overview of the relationship of use of information source and improvement in adjustment (RMSE)

In terms of information used, we can from figure 7.10 and figure 7.11, which can be seen that people with a technical background are more inclined to use various types of information when interacting, and relying on personal judgment and model information, historical data analysis would reduce the prediction error more. People without a technical background seem unable to make good use of various types of information, the information they rely on is more single, and in terms of technical information and personal judgment, they cannot interact well with the model information, so that their changes all make the forecasting error larger.

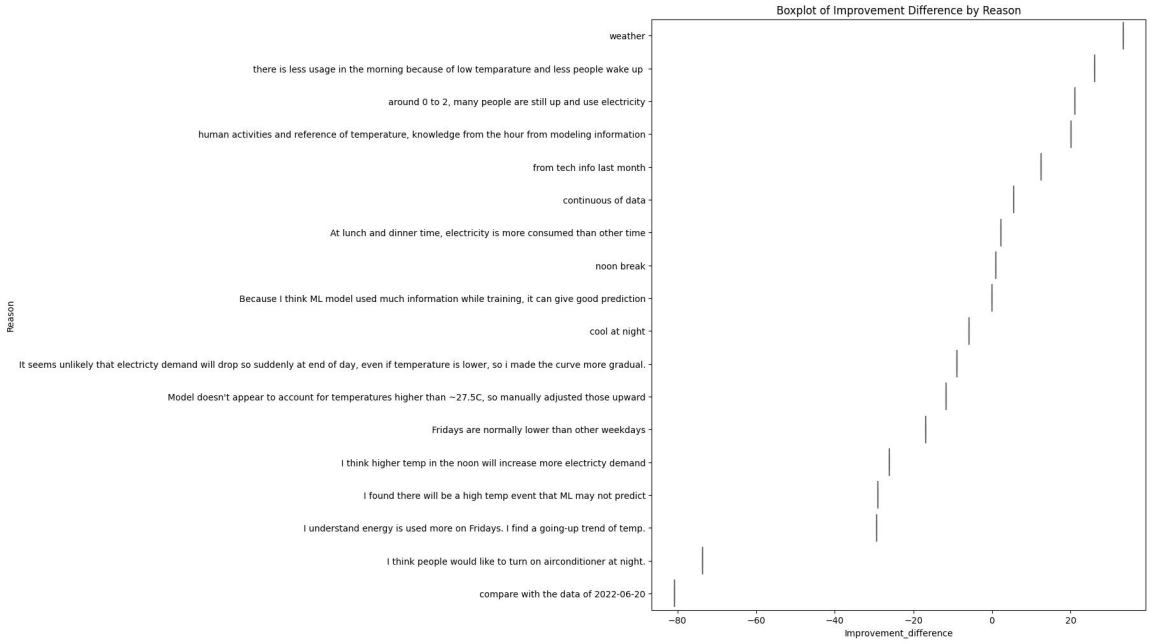

Figure 7.12 Boxplots showing the distribution of reasons for making adjustments and RMSE improvements from adjustments

The study also shows the reasons participants gave for making adjustments and the corresponding RMSE improvement. Though no strong patterns emerged, a

simple analysis reveals that different adjustment reasons may lead to varying degrees of change in RMSE. However, it can be superficially observed through this study that adjustments citing more specific and objective information tend to yield lower RMSE, such as “compare with data of 2020-06-20” and “I found there will be a high temp event that ML may not predict”.

Besides the survey after each adjustment, the dataset also records participants' behavioral logs. After processing, the dataset contains users' interactive behaviors with the system. Each row records one user behavior, including the time of occurrence, behavior type, performing user, and user group.

The columns include:

Column Name	Description
time	Time of behavior occurrence
behaviours	Behavior type, e.g. "start experiment", "view technique information", "interaction with info".
Name	Name of performing user
user_group	User group, either "group 1" or "group 2".

Table 7.3 Table of user behaviour logs

time	behaviours	Name	user_group
2023-06-28 11:29:51	start experiment	User_1	group 2
2023-06-28 11:30:11	view technique information	User_1	group 2
2023-06-28 11:30:42	interaction with info	User_1	group 2
2023-06-28 11:31:29	view technique information	User_1	group 2
2023-06-28 11:31:49	interaction with info	User_1	group 2

Figure 7.13 Top 5 rows of behaviour log dataset

This dataset as shown in figure 7.13, provides a comprehensive log of user interactions within an interactive dashboard designed for experts to revise electricity demand predictions. Each row in the dataset represents a distinct user action, characterized by four attributes. The 'time' column records the timestamp of the action in the "YYYY-MM-DD HH:MM:SS" format, offering a chronological context to the user behavior. The 'behaviours' column categorizes the nature of the user interaction, such as the start of an experiment, viewing technique information, or other forms of interaction with the system. The 'Name' column identifies the user performing the action, represented by labels like 'User_1', 'User_2', etc. Lastly, the 'user_group' column denotes the group membership of the user, allowing for comparative analysis between different user cohorts. Together, these elements provide valuable insights into the user engagement patterns, preferences, and behaviors within the system, thereby facilitating a user-centered optimization of the prediction model and the interface design.

Behaviour	Explanation
'make global change'	User made a global adjustment, which affects the entire demand curve. This could be due to the user identifying a general trend that they believe the machine learning model has not captured.
'view technique information'	User viewed information about the prediction technique or model. This suggests that users are interested in understanding how the predictions were generated when making their adjustments.
'interaction with info'	User interacted with the information provided in the system. This could include actions such as reading news articles or exploring feature data.
'make local change'	User made a local adjustment, affecting only a specific portion of the demand curve. This could be due to the user identifying a specific area where they believe the machine learning model has erred.
'view contextual information'	User viewed contextual information provided in the system. This could include actions such as reading news articles or exploring feature data.
'feedback'	User provided feedback on the system or their experience. This could be through a survey or other feedback mechanism in the system.
'confirm change'	User confirmed their adjustments, indicating they are satisfied with the changes made.
'view modeling information'	User viewed information related to the modeling process or features used in the prediction.
'start experiment'	User started a new session or experiment with the system.
'End experiment'	User ended a session or experiment with the system.
'undo'	User undid a previous adjustment, indicating they were not satisfied with the change.
'reset plot'	User reset all adjustments, discarding all changes they made to the demand curve. This suggests the user wanted to start over with their adjustments.

Table 7.4 Types of behaviours collected from experiment and explanations

This table in table 7.4 provides a detailed overview of the different user behaviors recorded in the log data, each corresponding to a specific interaction within the system. Behaviors range from making global or local adjustments to the demand curve, to interacting with various informational resources, to administrative actions such as starting or ending an experiment. For instance,

'make global change' represents instances where users adjust the entire demand curve, likely in response to perceived overarching trends in the data. 'View technique information' and 'view modeling information' indicate users' engagement with the underlying prediction methodologies, suggesting an interest in understanding the mechanics of the forecasts. Actions like 'confirm change', 'undo', and 'reset plot' reflect users' iterative refinement of their adjustments, indicating the dynamic nature of their engagement with the system. Feedback mechanisms are also logged, capturing users' responses to their experience. By examining the frequency and context of these behaviors, we can gain valuable insights into users' strategies, preferences, and the utility of different system features.

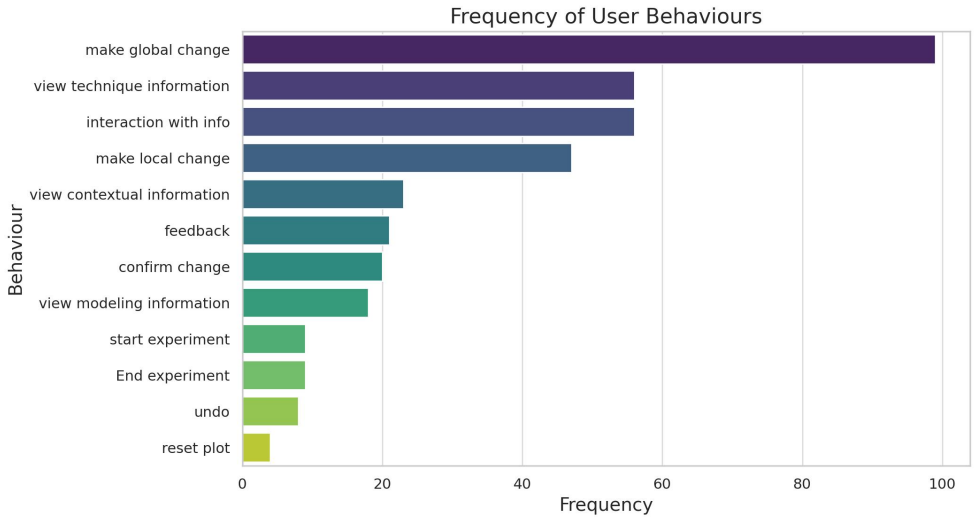

Figure 7.14 Frequency of user behaviours

The bar chart in figure 7.14, visualizes the frequency of different user behaviors logged in the system. Each bar represents a distinct type of user behavior, and the length of the bar corresponds to the frequency of that behavior. From the chart, we

can see that 'make global change' is the most frequent behavior, which suggests that users often prefer to make adjustments that affect the entire demand curve. This could be due to users identifying overall trends in the data that they believe the machine learning model has not adequately captured. On the other hand, 'reset plot' is the least frequent behavior. This suggests that users rarely choose to discard all their adjustments and start over, indicating a general level of confidence in the changes they have made. Behaviors such as 'view technique information', 'interaction with info', and 'make local change' are moderately frequent, suggesting that users actively engage with the information provided and make targeted adjustments based on this information.

By understanding the behaviors log data, the visualization of users' operational processes can be achieved, and the operational processes also imply users' decision-making processes. In the previous subsection, the results mentioned that participants with technical backgrounds made better adjustments compared to those without technical backgrounds in this study. Compared to previous studies (Sanders & Ritzman, 1992), by collecting users' behavior logs of interacting with the user interface, this study can not only give an answer, but also understand why by delving into the interaction process to discover and uncover differences in the operational patterns of users with different backgrounds, and attempt to find some regularities. Moreover, by drawing the This task represents an electricity market prediction task but is fundamentally a decision-making task for the individual. From a task, goal, necessary information to complete the task, and the decision to adjust (and its range/intensity), how people complete the task is traced. Groups 1 and 2 are mainly compared.

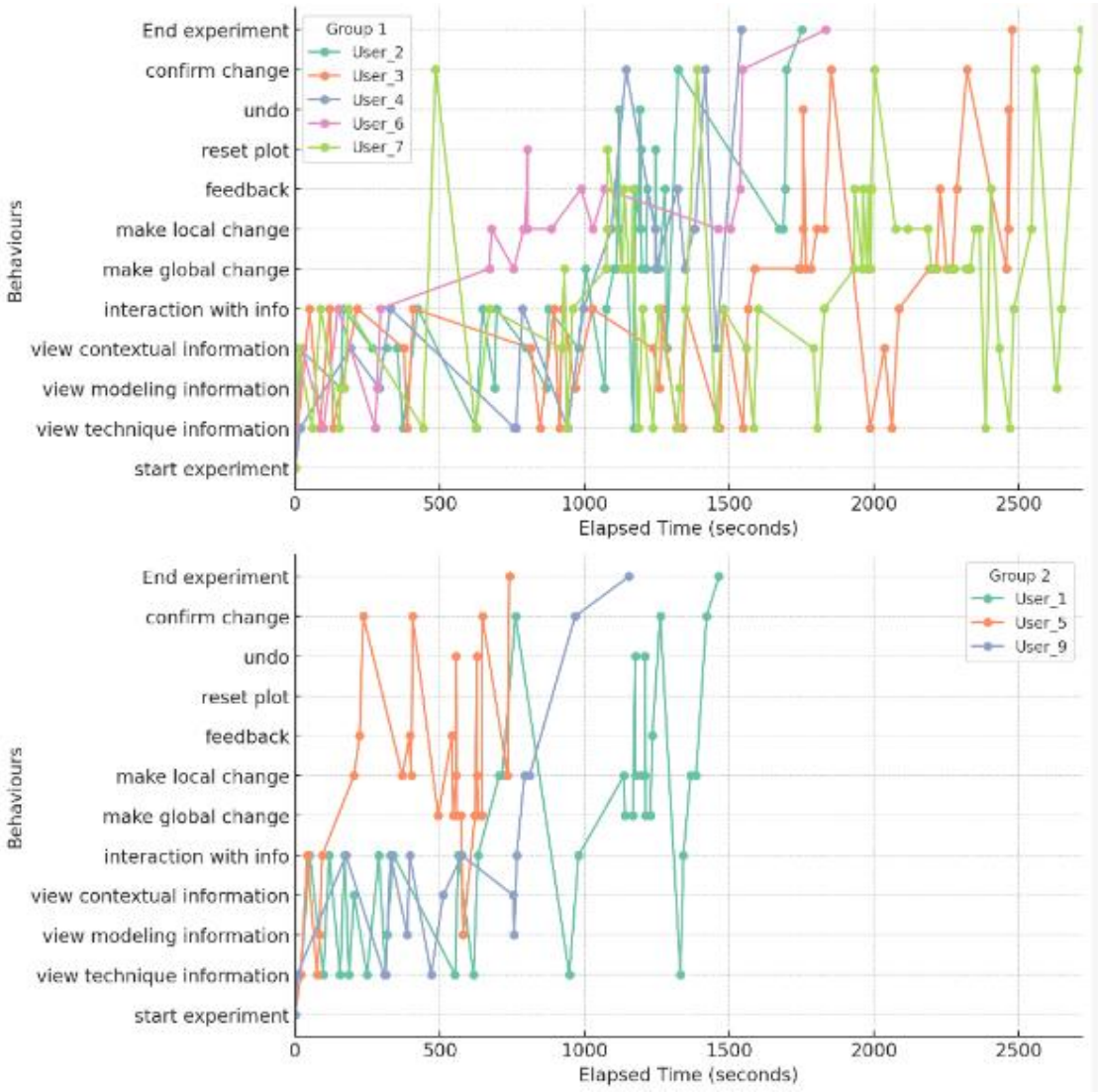

Figure 7.15 Comparison between the behaviour log sequences of group 1 and group 2

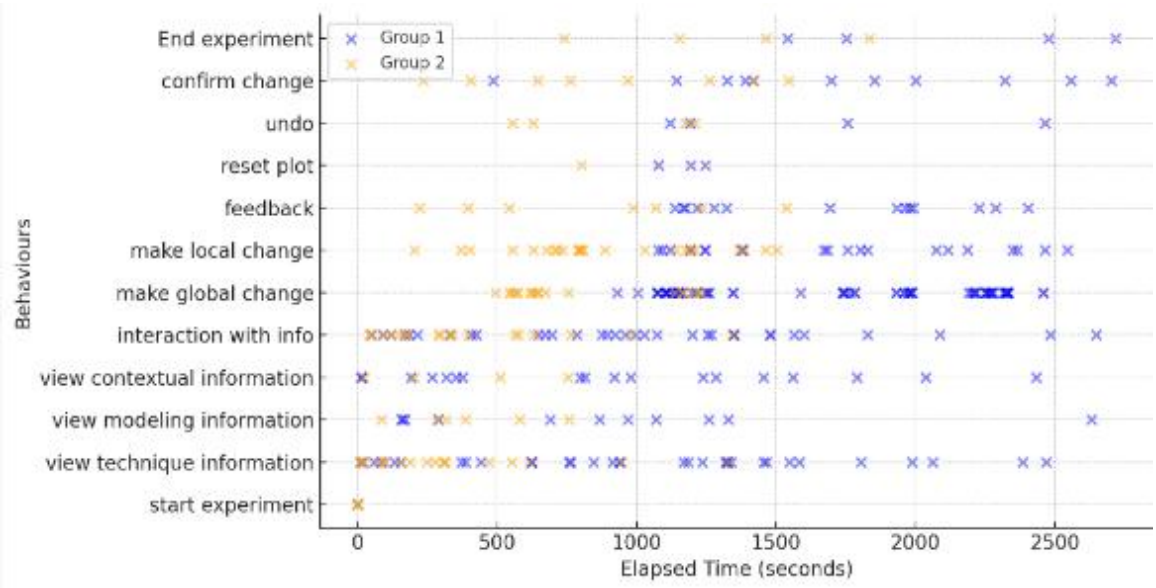

Figure 7.16 Comparison between the behaviour logs of group 1 and group 2 (scatter pot)

Through, examining overall procedural sequences from figure 7.15 and figure 7.16 reveal significant differences. Group 1 spent more time completing the task and adjusted more frequently. In the initial information reception stage, they spent longer receiving information and interacted with information more frequently. Group 2 behaved more simply, tending to promptly adjust machine forecasts and iterate a few adjustments to complete the experiment. Group 1 behaviors were richer: after understanding information and formulating a plan, they made multiple adjustments; those adjusting repeatedly constantly interacted with information to confirm their steps. Overall, Group 1 knew from the start how to accomplish goals through AI, what information to acquire, and obtained information more richly before making extensive adjustments while constantly interacting with information to validate and optimize each round. Without technical backgrounds, Group 2 did not know the initial or subsequent adjustment approaches, thus adjusting rapidly through behavior to complete the experiment. This suggests that without technical

backgrounds, individuals may not know the steps they should initially take for unfamiliar tasks.

By studying the behavior logs, this research gained a deeper understanding and examined the differences in operational processes between groups with and without technical backgrounds. The operational processes to some extent also represent the users' decision-making processes. As shown in Figure 7.17, after initially recognizing the problem, users would collect information, corresponding to the review of information stage in our research. Then users need to evaluate their decisions to identify options and judge the range and intensity of their decisions, finally making a decision, which may receive feedback depending on the degree of options. The findings of this study not only proved that people with technical backgrounds outperformed those without technical backgrounds in results, but also, by collecting user interface interaction data, delved deeper under the results and pointed out that for people with technical backgrounds, they could obtain desired information more efficiently and actively, which we refer to as active information seekers. In contrast, people without technical backgrounds cannot discern the next steps when interacting with machines, thus their information acquisition is more limited, and they may even be unable to discern the value implied in the information. We refer to them as passive information receivers. These two types exhibit different behaviors when interacting with machines. People with technical backgrounds can acquire information more frequently and deeply, and make adjustments afterwards. As people without technical backgrounds cannot obtain sufficient information for judgment, they enter the adjustment stage more quickly and complete the experiment rapidly. These findings provide some value in the direction of human-AI collaboration in supporting decision-making.

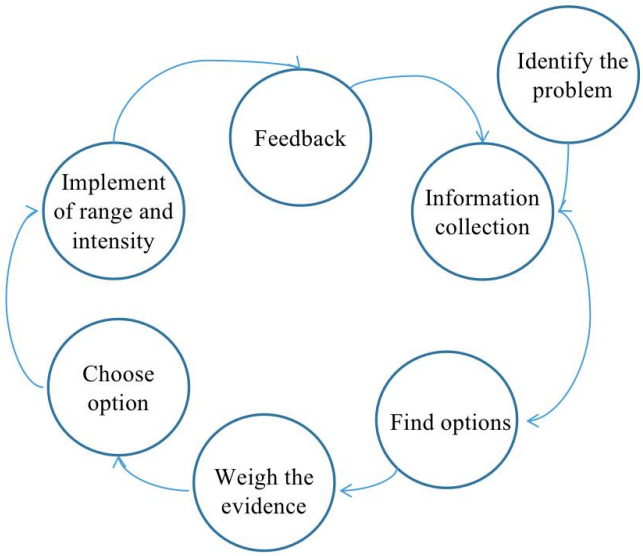

Figure 7.17 Human decision flow

Subsequently, this research proposes a transition plot that can visualize and trace users' behaviors on the user interface in a more in-depth manner. Through the novel framework proposed in this study, the following can be achieved:

- The operational process of humans in human-AI collaboration can be understood more quickly and directly.
- By incorporating order as shown in Figure 7.18, or transition probabilities as shown in Figure 7.19 into the transition matrix, users' preferences in decision-making can be discerned to some extent. Through this coding, the roles played by humans in interactions can be better understood.

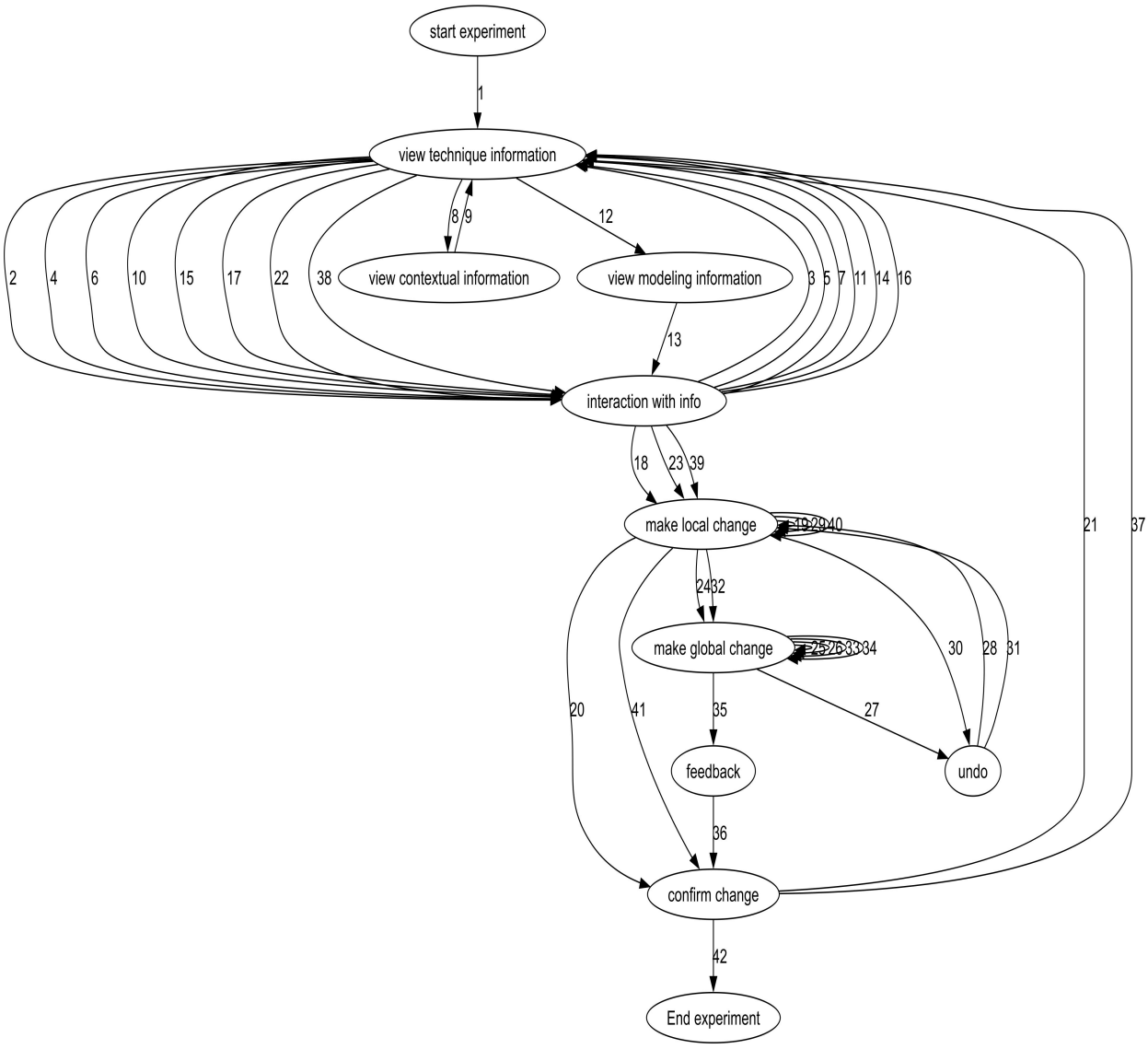

Figure 7.18 Visualization of operational process with order indicators

Figure 7.18 shows the operational process of one user. Through this operational process, it can be observed that the user prefers to acquire information by viewing technical information, while rarely interacting with contextual information. There is no salient distinction exhibited between making local changes and global changes. In the adjustment stage, this user only utilized “feedback” once and “uno” multiple times, potentially demonstrating this user's high confidence in their own

adjustments. By analyzing this visualization, the experiences and some preferences of this user when interacting with the user interface can be understood very clearly and simply.

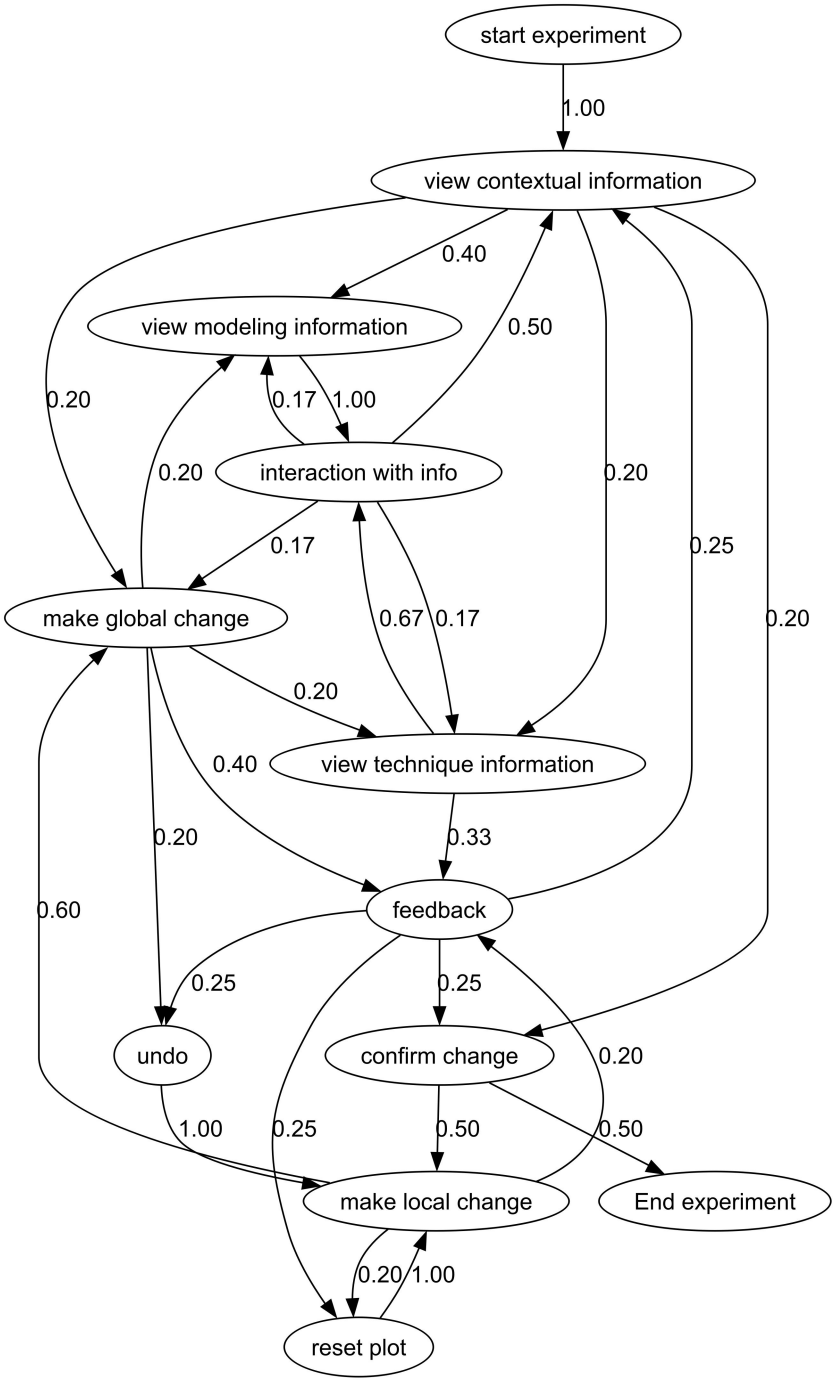

Figure 7.19 Visualization of operational process with probability transition

Figure 7.19 visualizes the operational process of another participant with probability transitions incorporated. From the visualization, it can be observed that this participant first understands the contextual information, and then has a higher probability of acquiring information from the modeling information. This user has a higher probability of interacting with the information after reviewing it. This demonstrates that the user has a strong technical background and a clear cognitive plan for the overall actions. Although the framework and visualization proposed in this study are not yet comprehensive, many relevant and insightful information can still be obtained from this framework and visualization.

In summary, this study investigated whether human adjustments could improve machine learning predictions of electricity demand, and whether technical background influenced adjustment accuracy. The results showed mixed evidence regarding humans improving machine forecasts overall. While one statistical test found no significant difference, limiting analysis to times when machine predictions faltered revealed a benefit of human adjustment.

Further analysis found those with technical backgrounds consistently made more accurate adjustments than non-technical participants, even significantly so during periods of poor machine forecasting. Visualizing and analyzing user interaction logs uncovered meaningful behavioral differences between groups. Those with technical expertise actively sought information to understand the task before making incremental changes, while non-technical users quickly made adjustments to finish the task.

Although sample sizes were small, results suggest technical knowledge allows humans to complement AI in beneficial ways. When machine learning falters, human expertise can detect issues and improve forecasts. Interaction trace visualizations also showcase how technical users leverage diverse information to

actively enhance their decisions. Future work should further investigate behavioral differences in human-AI collaboration across user groups and task contexts. Larger samples could solidify the impact of technical expertise. Overall, this research provides initial evidence that humans with relevant knowledge can meaningfully collaborate with AI systems for improved forecasting outcomes.

This chapter presented an experimental study examining how human expertise can improve machine learning forecasting accuracy. Through an interactive interface, technical and non-technical participants adjusted statistical electricity demand predictions. Results demonstrated mixed evidence that human adjustments enhanced overall machine forecasting. However, during periods of poor model performance, human input significantly improved accuracy. Tests also found technical users outperformed non-technical participants, especially when machine learning faltered. Analyzing interaction logs revealed technical users actively sought information to incrementally refine adjustments, while non-technical ones rapidly finished the task. Although limited by sample size, the experiments provide initial evidence that human cognition can meaningfully complement AI limitations if users have relevant knowledge. When statistical models struggle, human judgment can detect issues and enhance predictions. Further research should investigate these dynamics across contexts with larger samples. But overall, this work suggests integrated human-AI collaboration can yield improved forecasting compared to AI alone.

Chapter 8 - Conclusion

Chapter 8.1 Conclusion

This thesis has presented an in-depth examination of leveraging human-AI collaboration to improve electricity demand forecasting accuracy and reliability in Japan. Through a comprehensive literature review, the challenges and complexities of electricity demand forecasting were highlighted, including the impacts of extreme weather, intermittent renewable sources, evolving policies, and social factors unique to the Japanese context. The limitations of current machine-oriented approaches were also discussed, presenting opportunities for integrating human expertise to address data constraints, model interpretability issues, and modeling of rare events.

An innovative experimental framework was proposed and implemented to evaluate a collaborative forecasting approach incorporating both machine predictions and human adjustments. The results demonstrated the benefits of human input in enhancing model accuracy when statistical forecasts alone falter, such as during peak demand periods. Further analysis found that users with technical backgrounds consistently produced more accurate adjustments compared to non-technical participants. Examining user interaction logs also revealed meaningful process differences between these groups, with technical users actively seeking diverse information to incrementally refine their changes.

Chapter 8.2 Limitations and future works

While this thesis makes valuable contributions, certain limitations should be acknowledged. The sample size of participants in the human-AI collaborative forecasting experiment was relatively small, constraining the generalizability of the results. Additionally, the participant pool lacked diversity, with all users being

students at the same university. Broader samples could reveal different behavioral patterns and performance outcomes. The study was also limited to a simulated forecasting task rather than real-world energy demand predictions, which may impact user motivations and system dynamics. Furthermore, the proposed visualization frameworks require additional validation to demonstrate their utility in understanding human-AI interactions. Users' subjective survey responses and limited behavior data also pose measurement challenges. Going forward, larger-scale experiments, more diverse participants, real-world forecasting tasks, enhanced visualization techniques, and more objective user measurements could strengthen the research. Overall, this thesis provides important foundational insights, but continued work is needed to consolidate findings and address these limitations. The initial evidence for human-AI collaborative forecasting is promising, but further research can build on these limitations to expand the understanding of optimal integration approaches.

In terms of future directions, one area worthy of further exploration is the development of more comprehensive human-AI collaboration interfaces. This could involve functionality to confirm user intentions to adjust forecasts before changes are made. More seamless participation mechanisms can strengthen the human-AI partnership. Additionally, collecting more granular user behavior data - including adjustment magnitudes, durations, sequences, and other attributes - would enable deeper analysis of user operational and decision-making processes. Such insights can uncover optimal collaboration strategies. Reducing biases is another key goal, achievable through providing users greater environmental context and running concurrent experiments. Finally, modeling patterns in user behaviors through probabilistic and machine learning approaches could support users lacking technical expertise by predicting productive next steps.

Overall, this research highlights the potential for hybrid intelligence to produce more accurate, robust, and transparent energy demand forecasts through purposeful integration. The implications are far-reaching in the context of supporting complex power grid operations and planning in Japan and beyond. Further work on optimizing human-AI collaboration promises to unlock even greater capabilities. By respecting complementary human and machine strengths, forecasting systems can be enhanced to address multifaceted real-world challenges.

References or Bibliography

1. Abedin, B., Meske, C., Junglas, I. A., Rabhi, F. A., Motahari-Nezhad, H. R. (2022). Designing and Managing Human-ai Interactions. Inf Syst Front, 3(24), 691-697.
2. Abramson, B., & Clemen, R. T. (1995). Probability forecasting. International Journal of Forecasting, 11(1), 1-4.
3. Adya, M., & Lusk, E. J. (2012). Designing effective forecasting decision support systems: Aligning task complexity and technology support. Omega, 61, 196-211.
4. Ageng, D. K., Huang, C., Cheng, R. (2021). A short-term household load forecasting framework using LSTM and data preparation. IEEE Access, (9), 167911-167919.
5. Akimoto, K., Sano, F., Homma, T., Oda, J., Nagashima, M., & Kii, M. (2010). Estimates of GHG emission reduction potential by energy-saving and low-carbon technologies in the residential sector. Applied Energy, 87(9), 2790-2799.
6. Alharbi, F. R., Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average With Exogenous Factors (Sarimax) Forecasting Model-based Time Series Approach. Inventions, 4(7), 94.
7. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
8. Amato, U., Antoniadis, A., Feis, I. D., Goude, Y., Lagache, A. (2021). Forecasting High Resolution Electricity Demand Data With Additive Models Including Smooth and Jagged Components. International Journal of Forecasting, 1(37), 171-185.

9. Amjady, N., & Keynia, F. (2009). Short-term load forecasting of power systems by combination of wavelet transform and neuro-evolutionary algorithm. Energy, 34(1), 46-57.
10. Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., Mané, D. (2016). Concrete problems in AI safety.
11. Armstrong, J. S. (2001). Principles of forecasting: A handbook for researchers and practitioners.
12. Bao, Y., Cheng, X., Vreede, T. d., Vreede, G. d. (2021). Investigating the Relationship Between Ai And Trust In Human-ai Collaboration. Proceedings of the Annual Hawaii International Conference on System Sciences.
13. Bengio, Y., Courville, A., Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(35), 1798-1828.
14. Bergmeir, C. and Benítez, J.M. (2012) On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, pp.192-213.
15. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70-83.
16. Boulesteix, A., & Schmid, M. (2014). Machine learning versus statistical modeling. Biom. J., 56(4), 588-593.
17. Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting.
18. Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. AI Magazine, 26(4), 53-60.
19. Burton, J. W., Stein, M., Jensen, T. B. (2019). A Systematic Review Of Algorithm Aversion In Augmented Decision Making. J Behav Dec Making, 2(33), 220-239.

20. Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233-234.
21. Cancelo, J. R., Espasa, A., & Grafe, R. (2008). Forecasting the electricity
22. Cancelo, J. R., Espasa, A., & Grafe, R. (2008). Forecasting the electricity load from one day to one week ahead for the Spanish system operator. International Journal of Forecasting, 24(4), 588-602.
23. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
24. Chicco, D., Warrens, M. J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, (7), e623.
25. Chicco, D., Warrens, M., Jurman, G. (2021). the coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. Peerj Computer Science, (7), e623.
26. Christoffersen, P. F., & Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. Journal of Applied Econometrics, 11(5), 561-571.
27. Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5(4), 559-583.
28. Copeland, B. J. (2022). Artificial intelligence. In Encyclopædia Britannica.
29. Daugherty, P. R., & Wilson, H. J. (2018). Human+ machine: Reimagining work in the age of AI.
30. Dawid, A.P. (1984) Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society: Series A (General), 147(2), pp.278-292.

31. Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., Ebel, P. (2019). The Future Of Human-ai Collaboration: a Taxonomy Of Design Knowledge For Hybrid Intelligence Systems. Proceedings of the Annual Hawaii International Conference on System Sciences.
32. Dietvorst, B. J., Simmons, J. P., Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science, 3(64), 1155-1170.
33. Dineva, K., Atanasova, T. (2020). Systematic Look At Machine Learning Algorithms: Advantages, Disadvantages and Practical Applications. SGEM International Multidisciplinary Scientific GeoConference EXPO Proceedings.
34. Ding, D., Zhang, M., Pan, X., Yang, M., He, X. (2019). Modeling Extreme Events In Time Series Prediction. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
35. Edmundson, R. H., Lawrence, M. J., & O'Connor, M. J. (1988). The use of non-time series information in sales forecasting: A case study. Journal of Forecasting, 7(3), 201-211.
36. Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. Energy, 35(2), 512-517.
37. Elamin, N., Fukushige, M. (2018). Modeling and Forecasting Hourly Electricity Demand By Sarimax With Interactions. Energy, (165), 257-268.
38. Entezari, A., Aslani, A., Zahedi, R., Noorollahi, Y. (2023). Artificial intelligence and machine learning in energy systems: A bibliographic perspective. Energy Strategy Reviews, (45), 101017.
39. European Parliamentary Research Service. (2020). A governance framework for algorithmic accountability and transparency.
40. Fahad, M., Arbab, N. (2014). Factor Affecting Short Term Load Forecasting. JOCET, 4(2), 305-309.

41. Fan, S., & Hyndman, R. J. (2010). Forecast short-term electricity demand using semi-parametric additive model.
42. Fan, S., Hyndman, R. J. (2012). Short-term Load Forecasting Based On a Semi-parametric Additive Model. IEEE Trans. Power Syst., 1(27), 134-141.
43. Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence.
44. Figoli, F. A., Rampino, L., Mattioli, F. (2022). Ai In the Design Process: Training The Human-ai Collaboration. DS 117: Proceedings of the 24th International Conference on Engineering and Product Design Education (E&PDE 2022), London S.
45. Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces, 37(6), 570-576.
46. Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 25(1), 3-23.
47. Fildes, R., Goodwin, P., Lawrence, M., Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 1(25), 3-23.
48. Frikha, M., Taouil, K., Fakhfakh, A., Derbel, F. (2022). Limitation Of Deep-learning Algorithm For Prediction Of Power Consumption. Itise 2022.
49. García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer.
50. Gautam, A., Singh, V. (2020). Parametric Versus Non-parametric Time Series Forecasting Methods: a Review. JESTR, 3(13), 165-171.

51. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
52. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
53. Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. International Journal of Forecasting, 16(1), 85-99.
54. Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time series forecasting. Omega, 22(6), 553-568.
55. Goodwin, P., Gönül, M. S., & Önkal, D. (2014). Antecedents and effects of trust in forecasting advice. International Journal of Forecasting, 30(2), 354-366.
56. Hama, M. (2022, August 3). LNG price spike causes energy crises in strapped Asian nations. Nikkei Asia.Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. Biometrical Journal, 60(3), 431-449.
57. Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: An evaluation. Management Science, 27(2), 115-138.
58. Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. International Journal of Forecasting, 32(3), 896-913.
59. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice.
60. Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: principles and practice. OTexts.
61. Hyndman, R.J. and Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia.
62. Härdle, W. K. (2004). Nonparametric and Semiparametric Models. Springer Series in Statistics.

63. Islam, A. K. M. N., Ahmed, S. I., Ahmed, S. H., Smolander, K. (2022). What Influences Algorithmic Decision-making? a Systematic Literature Review On Algorithm Aversion. Technological Forecasting and Social Change, (175), 121390.
64. Ivanov, D., Dolgui, A., & Sokolov, B. (2019). The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. International Journal of Production Research, 57(3), 829-846.
65. JEPIC. (2023). The electric power industry in Japan 2023.
66. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: Springer.
67. Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 4(61), 577-586.
68. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 230-243.
69. Jiang, P., Li, R., Lu, H., Zhang, X. (2019). Modeling Of Electricity Demand Forecast For Power System. Neural Comput & Applic, 11(32), 6857-6875.
70. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
71. Kahneman, D., & Riepe, M. W. (1998). Aspects of investor psychology: Beliefs, preferences, and biases investment advisors should know about. Journal of Portfolio Management, 24(4), 52-65.
72. Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25.

73. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.
74. Khafaf, N. A., Jalili, M., Sokolowski, P. (2019). Application Of Deep Learning Long Short-term Memory In Energy Demand Forecasting. Engineering Applications of Neural Networks, 31-42.
75. Khatoon, S., Ibraheem, A. K. Singh, P. (2014). Effects of various factors on electric load forecasting: An overview. In 2014 6th IEEE Power India International Conference (PIICON), (pp. 1-5). Delhi, India: IEEE.
76. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).
77. Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid, 10(1), 841-851.
78. Krstonijević, S. (2022). Adaptive Load Forecasting Methodology Based On Generalized Additive Model With Automatic Variable Selection. Sensors, 19(22), 7247.
79. Lai, Y., Kankanhalli, A., Ong, D. C. (2021). Human-ai Collaboration In Healthcare: a Review And Research Agenda. Proceedings of the Annual Hawaii International Conference on System Sciences.
80. Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. Management Science, 32(12), 1521-1532.
81. Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. Organizational Behavior and Human Decision Processes, 43(2), 172-187.

82. Lawrence, M., Edmundson, R.H, & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time series. International Journal of Forecasting, 1(1), 25-35.
83. Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. International Journal of Forecasting, 22(3), 493-518.
84. Lawrence, M., O'Connor, M., & Edmundson, R. (2000). A field study of sales forecasting accuracy and processes. European Journal of Operational Research, 122(1), 151-160.
85. Lee, J., Cho, Y. (2022). National-scale Electricity Peak Load Forecasting: Traditional, Machine Learning, or Hybrid Model?. Energy, (239), 122366.
86. Leitão, D., Saleiro, P., Figueiredo, M. A., & Bizarro, P. (2022). Human-AI collaboration in decision-making: beyond learning to defer. arXiv preprint arXiv:2206.13202.
87. Li, B., Lu, M., Zhang, Y., Huang, J. (2019). a Weekend load forecasting model based on semi-parametric Regression Analysis Considering Weather and load Interaction. Energies, 20(12), 3820.
88. Li, C., Lu, R. (2023). Short-term Power Forecasting Model Based On Gwo-lstm Network. J. Phys.: Conf. Ser., 1(2503), 012039.
89. Lim, J. S., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: Its effectiveness and biases. Journal of Behavioral Decision Making, 9(3), 149-168.
90. Liu, J., Chen, R., Liu, L., Harris, J. L. (2006). A Semi-parametric Time Series Approach In Modeling Hourly Electricity Loads. J. Forecast., 8(25), 537-559.
91. Liu, P. (2022). Time series forecasting based on ARIMA and LSTM. Proceedings of the 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022).

92. Lu, Z. (2010). The elements of statistical learning: Data mining, inference, and prediction. Journal of the Royal Statistical Society Series A (Statistics in Society), 173(3), 693-694.
93. Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 32(1), 77-108.
94. METI (Ministry of Economy, Trade and Industry) (2018). Strategic energy plan.
95. Mahmoud, H. (2021). Parametric Versus Semi and Nonparametric Regression Models. IJSP, 2(10), 90.
96. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), e0194889.
97. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods.
98. Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2018). Statistical and Machine Learning Forecasting Methods: Concerns And Ways Forward. PLoS ONE, 3(13), e0194889.
99. Mandal, P., Senjyu, T., & Funabashi, T. (2006). Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market. Energy Conversion and Management, 47(11-12), 2128-2142.
100. Mandal, P., Senjyu, T., Urasaki, N., & Funabashi, T. (2007). A neural network based several-hour-ahead electric load forecasting using similar days approach. International Journal of Electrical Power & Energy Systems, 29(8), 598-609.
101. Mariano-Hernández, D., Hernández-Callejo, L., Solís, M., Zorita-Lamadrid, A., Duque-Pérez, O., Gonzalez-Morales, L., ... & García, F. S. (2022).

Comparative Study Of Continuous Hourly Energy Consumption Forecasting Strategies With Small Data Sets To Support Demand Management Decisions In Buildings. Energy Science & Engineering, 12(10), 4694-4707.

102. Marino, D., Amarasinghe, K., Manic, M. (2016). Building Energy Load Forecasting Using Deep Neural Networks. IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society.

103. Marmier, F., Cheikhrouhou, N. (2010). Structuring and Integrating Human Knowledge In Demand Forecasting: A Judgemental Adjustment Approach. Production Planning & Control, 4(21), 399-412.

104. Mathews, B. P., & Diamantopoulos, A. (1986). Managerial intervention in forecasting. An empirical investigation of forecast manipulation. International Journal of Research in Marketing, 3(1), 3-10.

105. Mathews, B. P., & Diamantopoulos, A. (1990). Judgmental revision of sales forecasts: Effectiveness of forecast selection. Journal of Forecasting, 9(4), 407-415.

106. McCarthy, J. (2007). What is artificial intelligence? Stanford University.

107. McCorduck, P. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. A K Peters/CRC Press.

108. Meier, J., Schneider, S., Le, C., Schmidt, I. (2020). Short-term Electricity Price Forecasting: Deep Ann Vs Gam. Information and Communication Technologies in Education, Research, and Industrial Applications, 257-276.

109. Meier, J., Schneider, S.A., & Le, C. (2019). Short-term Electricity Price Forecasting Using Generalized Additive Models. ICTERI Workshops.

110. Mhlanga, D. (2023). Artificial intelligence and machine learning for energy consumption and production in emerging markets: A review. Energies, 2(16), 745.

111. Moss, L., Corsar, D., Shaw, M., Piper, I., Hawthorne, C. (2022). Demystifying the Black Box: The Importance Of Interpretability Of Predictive Models In Neurocritical Care. Neurocrit Care, S2(37), 185-191.
112. Nakashima, H. H., Mantovani, D. M. N., Junior, C. M. (2022). Users' Trust In Black-box Machine Learning Algorithms. REGE.
113. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.
114. Nilsson, N. J. (2010). The quest for artificial intelligence: A history of ideas and achievements. Cambridge University Press.
115. O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. International Journal of Forecasting, 9(2), 163-172.
116. Ohashi, H. (2010). Electric power industry reform in Japan. RIETI.
117. Oliva, R., & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. Production and Operations Management, 18(2), 138-151.
118. Onyema, E. M., Almuzaini, K., Onu, F. U., Verma, D., Gregory, U. S., Puttaramaiah, M. (2022). Prospects and Challenges Of Using Machine Learning For Academic Forecasting. Computational Intelligence and Neuroscience, (2022), 1-7.
119. Otsuka, A. (2019). Natural Disasters and Electricity Consumption Behavior: A Case Study Of The 2011 Great East Japan Earthquake. Asia-Pac J Reg Sci, 3(3), 887-910.
120. Otsuka, A., Haruna, S. (2016). Determinants Of Residential Electricity Demand: Evidence From Japan. IJESM, 4(10), 546-560.

121. Patrick, H., Monika, W., Max, S., Sebastian, V., Michael, V., Gerhard, S. (2023). Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction. Proceedings of the 28th International Conference on Intelligent User Interfaces.
122. Patro, S. G., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.
123. Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., ... & Ziel, F. (2022). Forecasting: Theory and Practice. International Journal of Forecasting, 3(38), 705-871.
124. Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J., Pardalos, P. M. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. International Journal of Production Research, 16(58), 4964-4979.
125. Qdr, Q. (2006). Benefits of demand response in electricity markets and recommendations for achieving them.
126. Raza, M. Q., & Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renewable and Sustainable Energy Reviews, 50, 1352-1372.
127. Raza, M. Q., Khosravi, A., & Nahavandi, S. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renewable and Sustainable Energy Reviews, 50, 1352-1372.
128. Razmerita, L., Brun, A. (2022). Collaboration In the Machine Age: Trustworthy Human-ai Collaboration. Learning and Analytics in Intelligent Systems, 333-356.
129. Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Antonelli, G., ... & Cherubini, A. (2022). Experimental Evidence Of Effective Human–ai Collaboration In Medical Decision-making. Sci Rep, 1(12).

130. Rosenkrantz, D. J., Vullikanti, A., Ravi, S. S., Stearns, R. E., Levin, S. A., Poor, H. V., ... & Marathe, M. V. (2022). Fundamental Limitations On Efficiently Forecasting Certain Epidemic Measures In Network Models. Proc. Natl. Acad. Sci. U.S.A., 4(119).

131. Russell, S., & Norvig, P. (2016). Artificial intelligence: A modern approach. Pearson.

132. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.

133. Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence (EUR 30117 EN). Publications Office of the European Union.

134. Sanders, N. R. (1992). Accuracy of judgmental forecasts: A comparison. Omega, 20(3), 353-364.

135. Sanders, N. R., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. Journal of Behavioral Decision Making, 5(1), 39-52.

136. Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. (2015). Effective judgmental forecasting in the context of fashion products. Journal of Operations Management, 36, 33-45.

137. Setiyorini, T., Frieyadie, F. (2020). Comparison Of Linear Regressions and Neural Networks For Forecasting Electricity Consumption. pilar, 2(16), 135-140.

138. Shah, I., Bibi, H. B., Ali, S., Wang, L., Yue, Z. X. (2020). Forecasting One-day-ahead Electricity Prices For Italian Electricity Market Using Parametric and Nonparametric Approaches. IEEE Access, (8), 123104-123113.

139. Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. (2013). A survey of forecast error measures. World Applied Sciences Journal, 24(24), 171-176.

140. Shinkawa, T. (2022, September 6). Japan‘s electricity market: Status and next steps forward [Presentation]. Electricity and Gas Market Surveillance Commission, Ministry of Economy, Trade and Industry.

141. Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems.

142. Tashman, L.J. (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting, 16(4), pp.437-450.

143. Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8), 799-805.

144. Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research, 204(1), 139-152.

145. Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on European data. IEEE Transactions on Power Systems, 22(4), 2213-2219.

146. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

147. Tiwari, T., Tiwari, T., Tiwari, S. (2018). How artificial intelligence, machine learning and deep learning are radically different?. IJARCSSE, 2(8), 1.

148. Trunk, A., Birkel, H., Hartmann, E. (2020). On the Current State of Combining human and artificial intelligence For Strategic organizational decision making. Bus Res, 3(13), 875-919.

149. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A. J. (2019). Machine Learning Algorithm Validation With a Limited Sample Size. PLoS ONE, 11(14), e0224365.
150. Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.
151. Vivoda, V. (2016). Energy Security In Japan.
152. Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing.
153. Vössing, M., Kühl, N., Lind, M., Satzger, G. (2022). Designing Transparency For Effective Human-ai Collaboration. Inf Syst Front, 3(24), 877-895.
154. Wang, P. (1995). On the working definition of intelligence. Center for Research on Concepts and Cognition, Indiana University.
155. Wang, P. (2019). On defining artificial intelligence. Journal of Artificial General Intelligence, 10(2), 1-37.
156. Weron, R., Misiorek, A. (2008). forecasting spot electricity prices: a comparison of parametric and semiparametric time series models. International Journal of Forecasting, 4(24), 744-763.
157. Willemain, T. R. (1991). The effect of graphical adjustment on forecast accuracy. International Journal of Forecasting, 7(2), 151-154.
158. Wood, S. N. (2006). Generalized additive models: An introduction with R. CRC press.
159. Wood, S. N. (2006). Generalized additive models: an introduction with R. CRC press.
160. Xu, P., Ji, X., Li, M., Lu, W. (2023). Small Data Machine Learning In Materials Science. npj Comput Mater, 1(9).

161. Yang, Y., Wu, J., Chen, Y., Li, C. (2013). A New Strategy For Short-term Load Forecasting. Abstract and Applied Analysis, (2013), 1-9.
162. Yuce, B., Mourshed, M., Rezgui, Y. (2017). A Smart Forecasting Approach To District Energy Management. Energies, 8(10), 1073.
163. Zhang, Q., Ishihara, K. N., Tezuka, T. (2012). Scenario Analysis On Future Electricity Supply and Demand In Japan. Energy, 1(38), 376-385.
164. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).
165. Russell, S. J., & Norvig, P. (2020). Artificial intelligence: a modern approach (4th ed.). Pearson.

Appendix A - Supporting materials

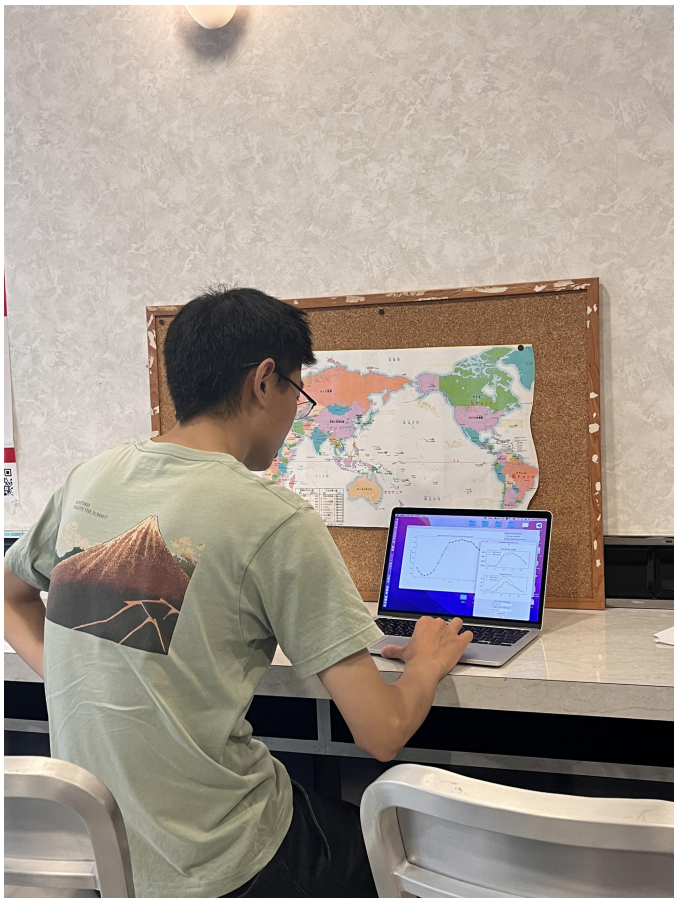

Image 1 Appendix A. In-progress experiment photo record

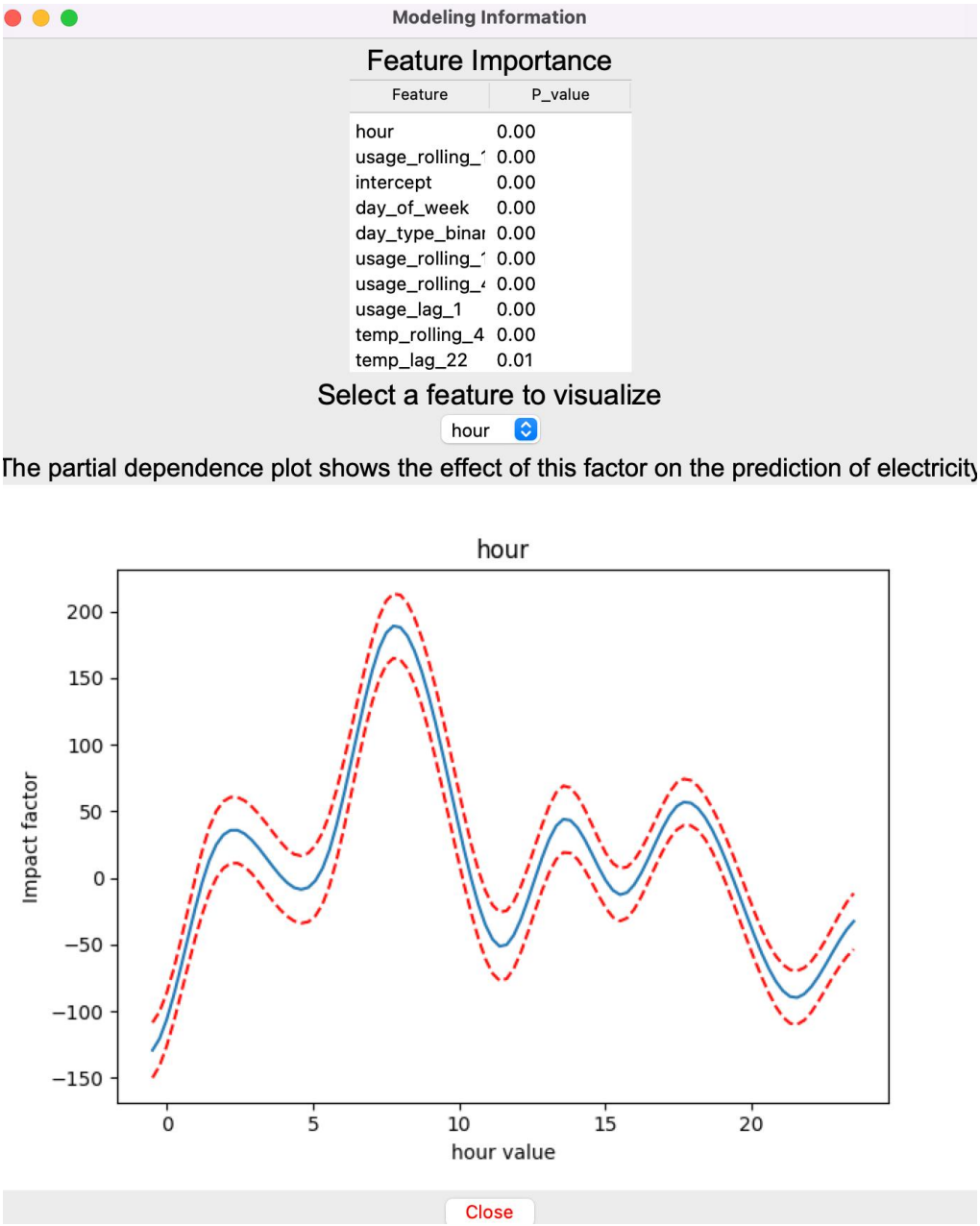

Figure 8.1 Appendix A. The user interface for modeling information

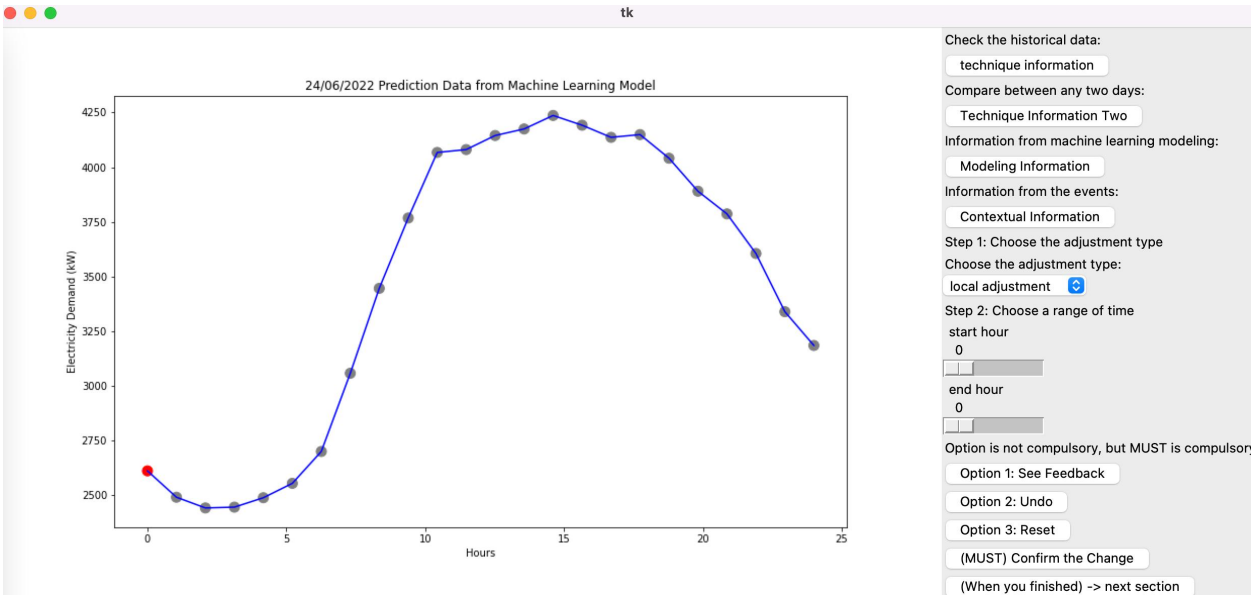

Figure 8.2 Appendix A. The home page of user interface

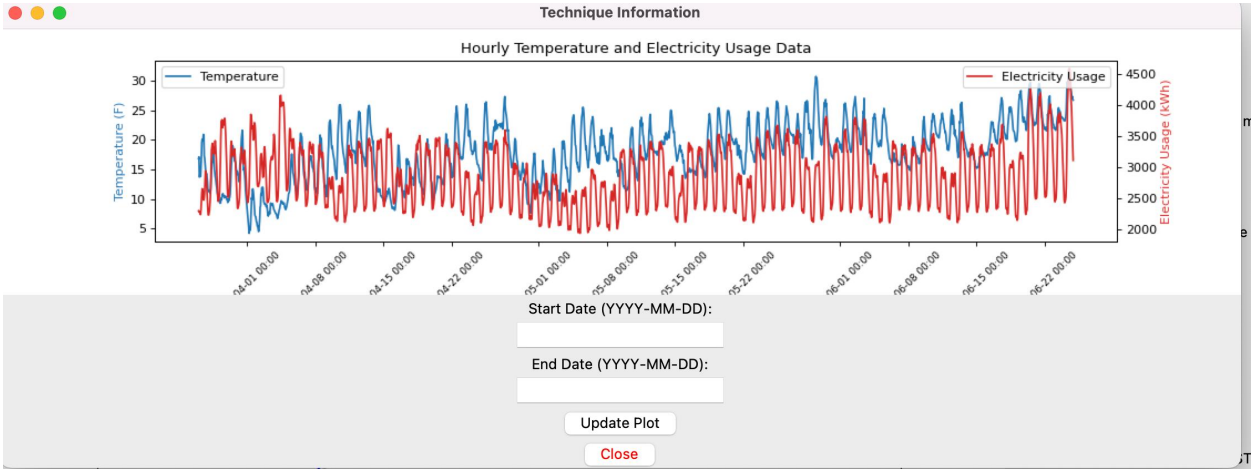

Figure 8.3 Appendix A. The user interface of historical data analysis

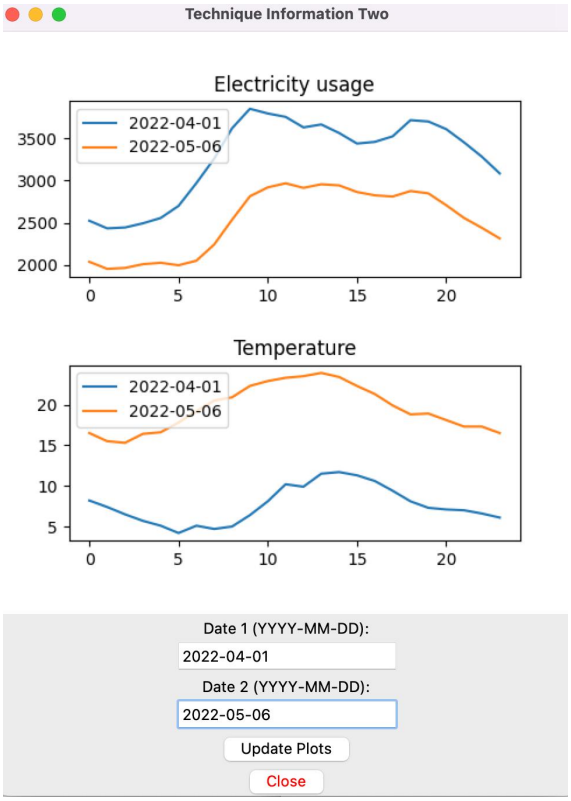

Figure 8.4 Appendix A. The user interface of historical data analysis (different date)

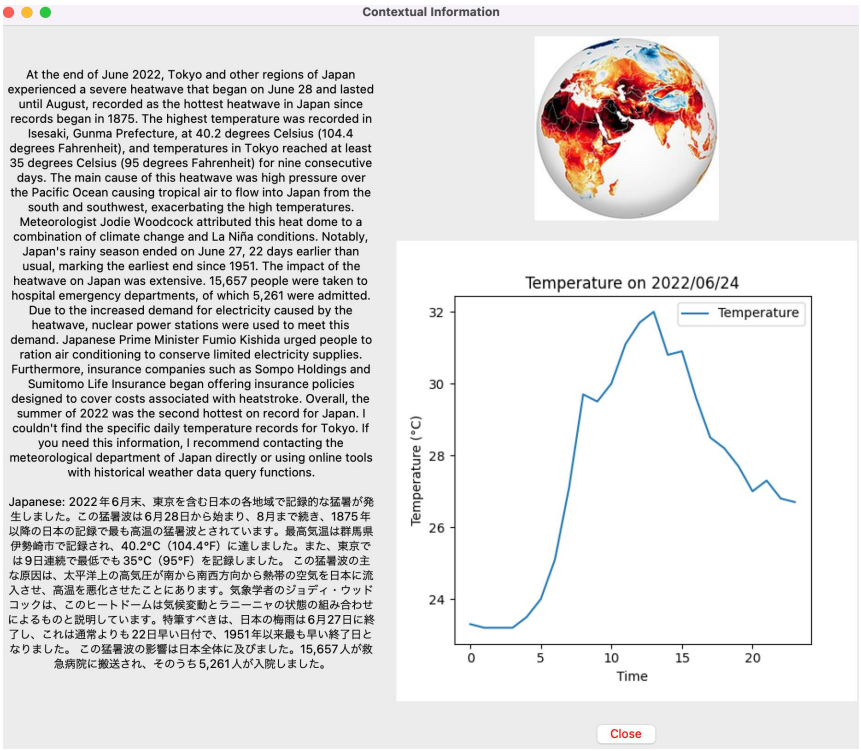

Figure 8.5 Appendix A. The user interface of contextual information

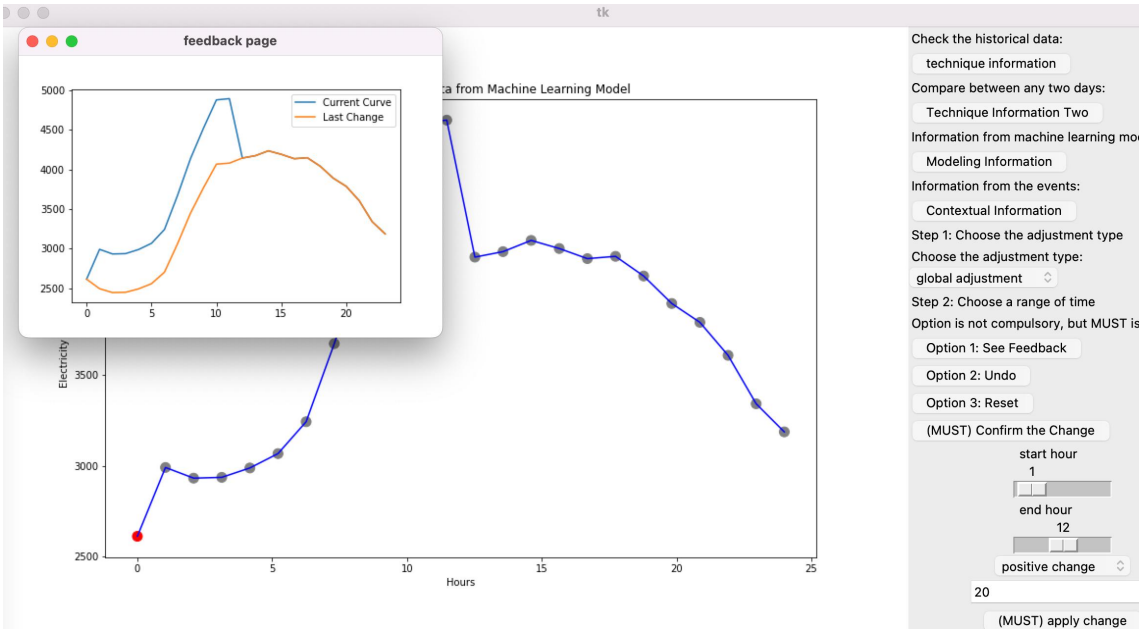

Figure 8.6 Appendix A. The user interface of feedback function

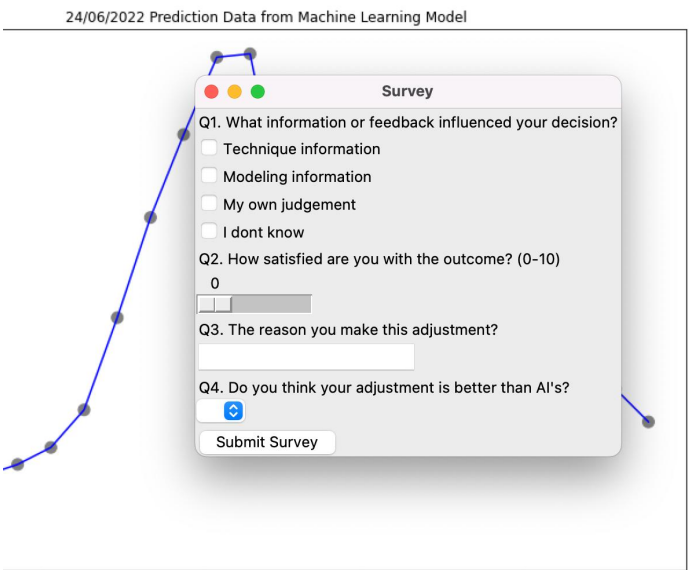

Figure 8.7 Appendix A. Short survey after each adjustment

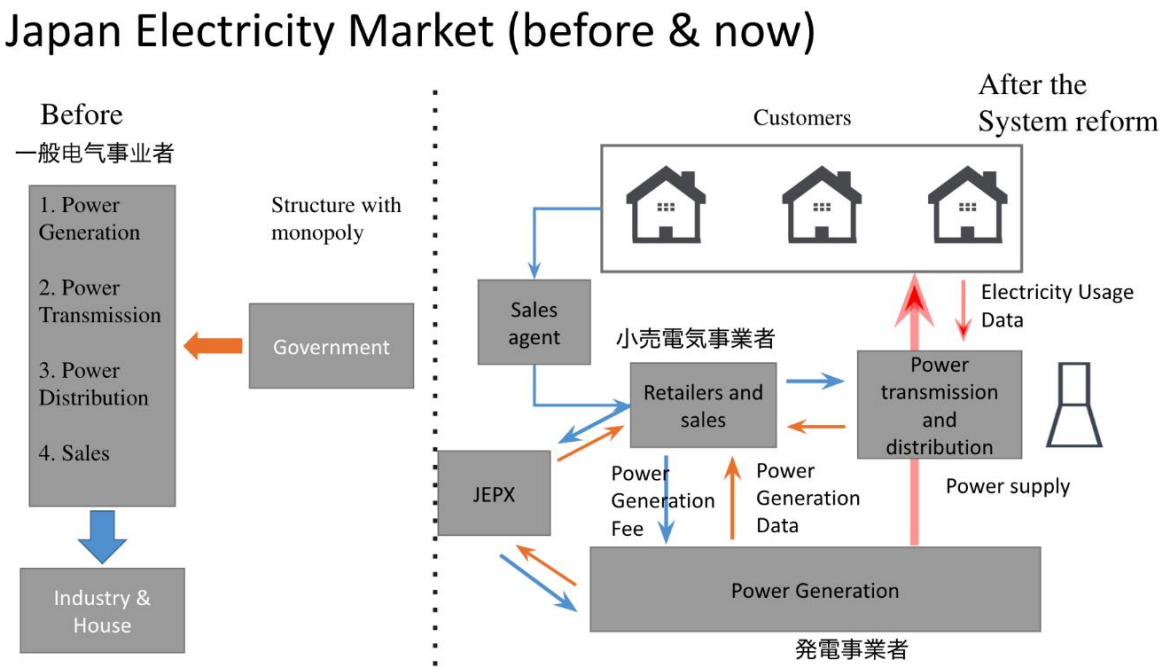

Figure 8.8 Appendix A. The market structure of Japan electricity market

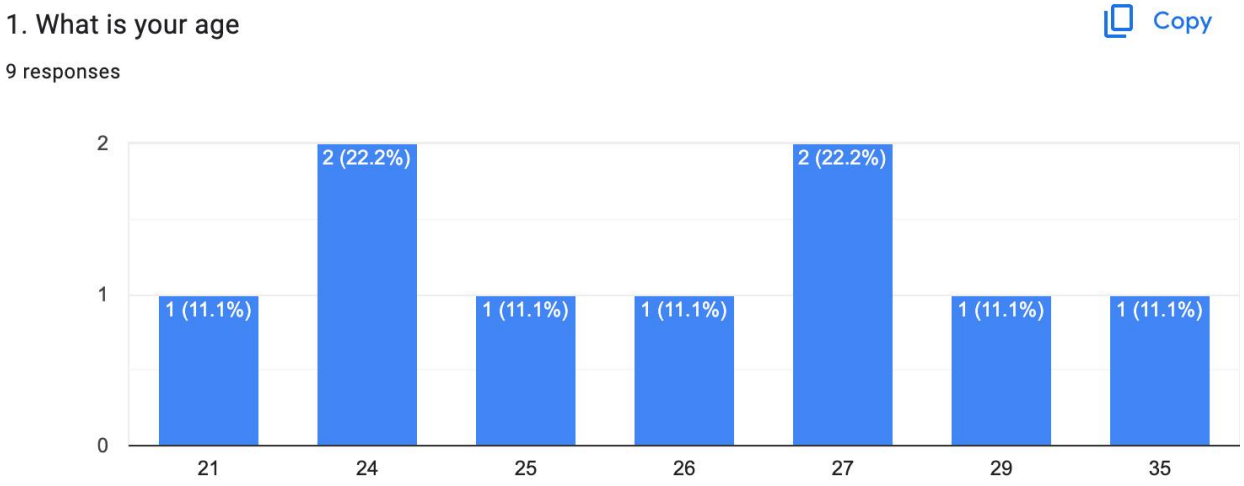

Figure 8.9 Appendix A. Survey results: age

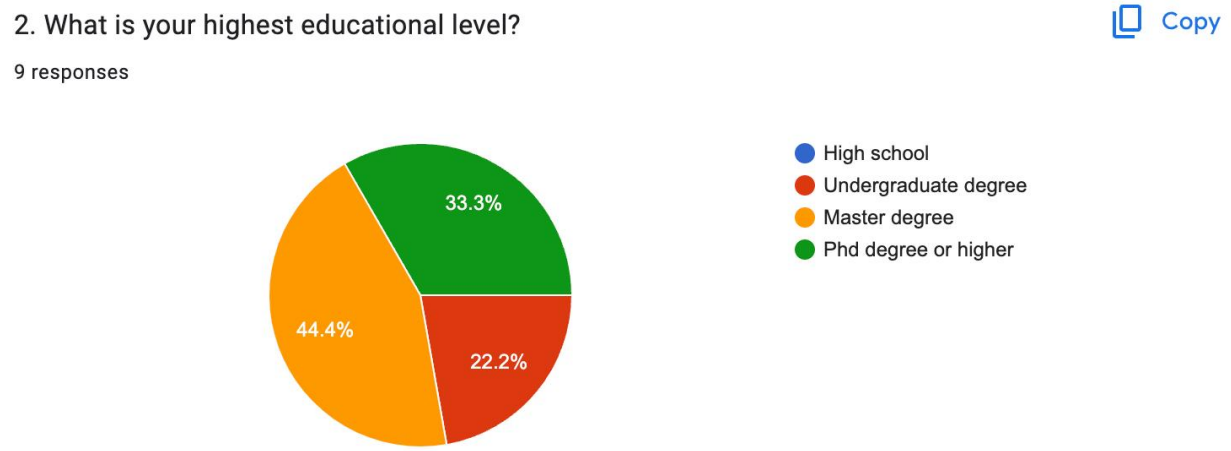

Figure 8.10 Appendix A. Survey results: education level

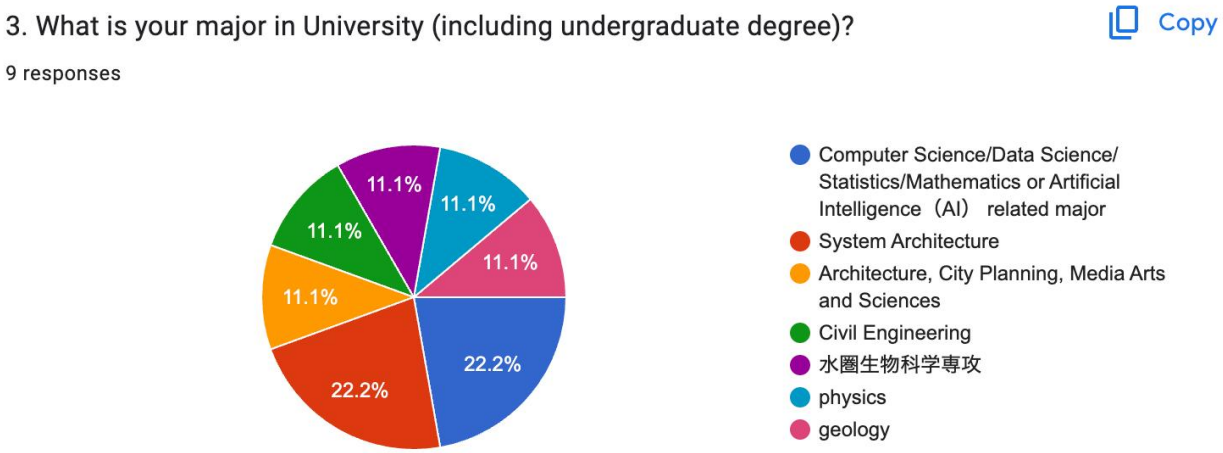

Figure 8.11 Appendix A. Survey results: major in university

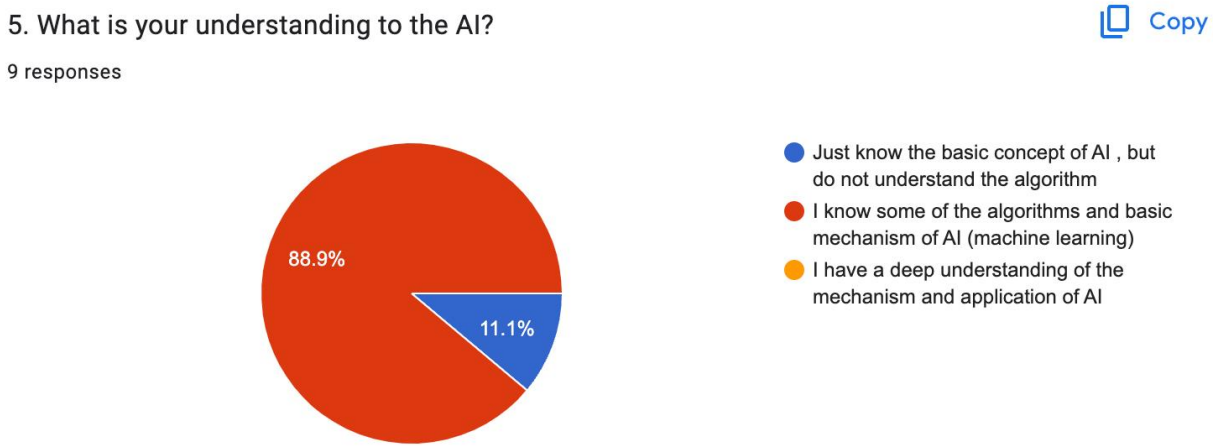

Figure 8.12 Appendix A. Survey results: understanding of AI

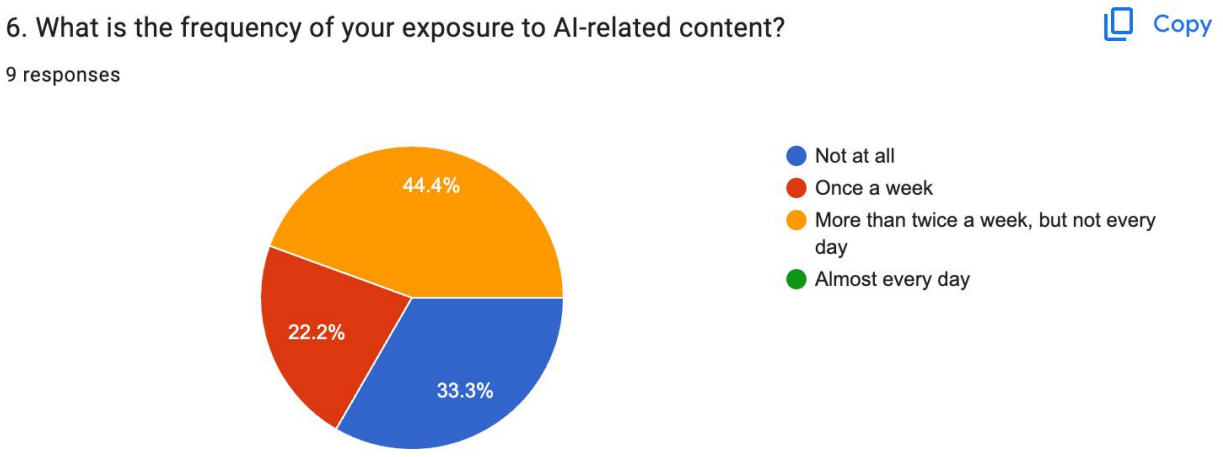

Figure 8.13 Appendix A. Survey results: frequency of exposure to AI-related content

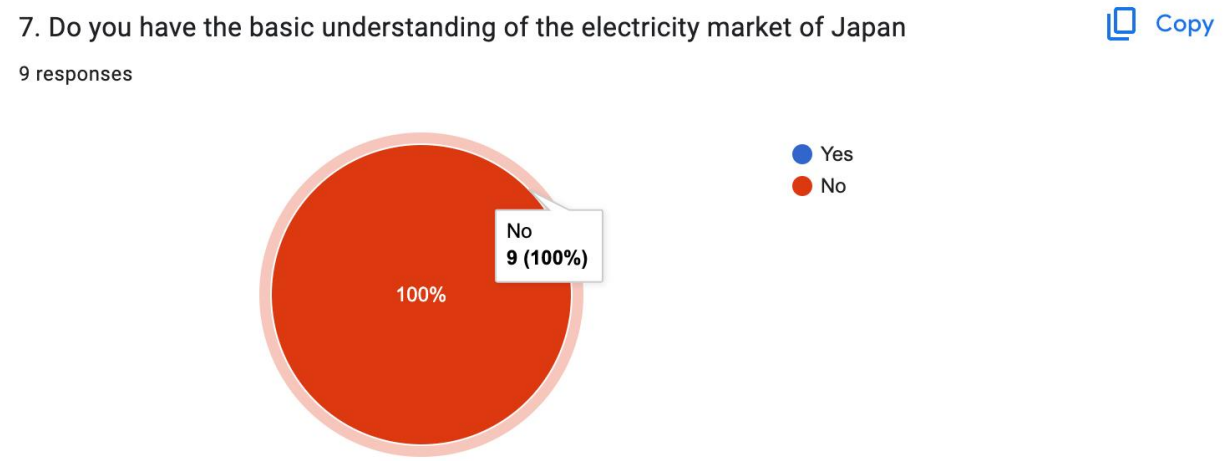

Figure 8.14 Appendix A. Survey results: understanding of electricity market of Japan

1. According to your judgment, how accurate will you rank the prediction result of AI? Copy

9 responses

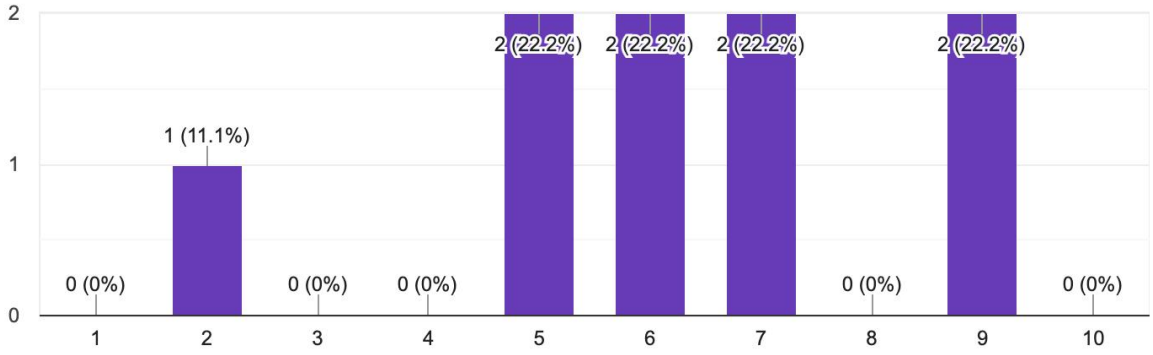

Figure 8.15 Appendix A. Survey results: rank of prediction accuracy of AI

2. According to your judgmental results, how accurate will you rank the prediction result of your adjusted results based on AI prediction? Copy

9 responses

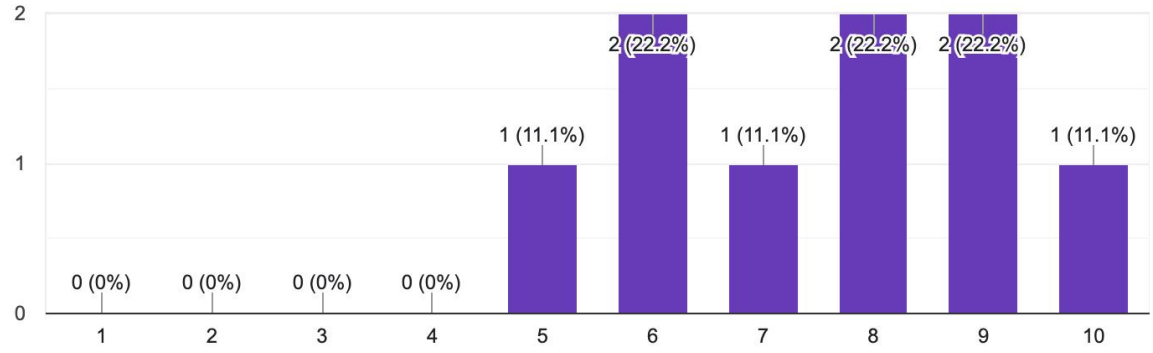

Figure 8.16 Appendix A. Survey results: rank of prediction accuracy of human adjustment

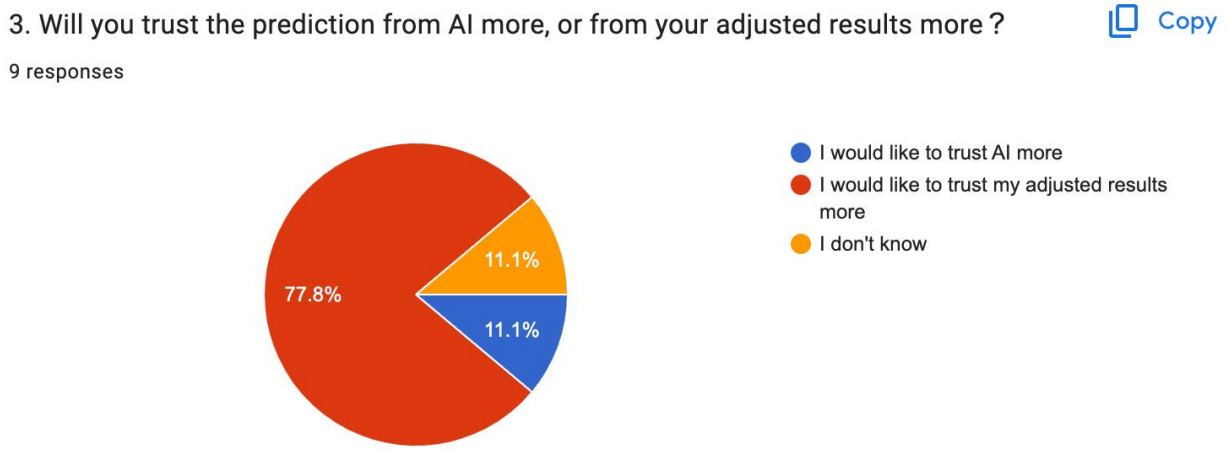

Figure 8.17 Appendix A. Survey results: trust between AI and human adjustments

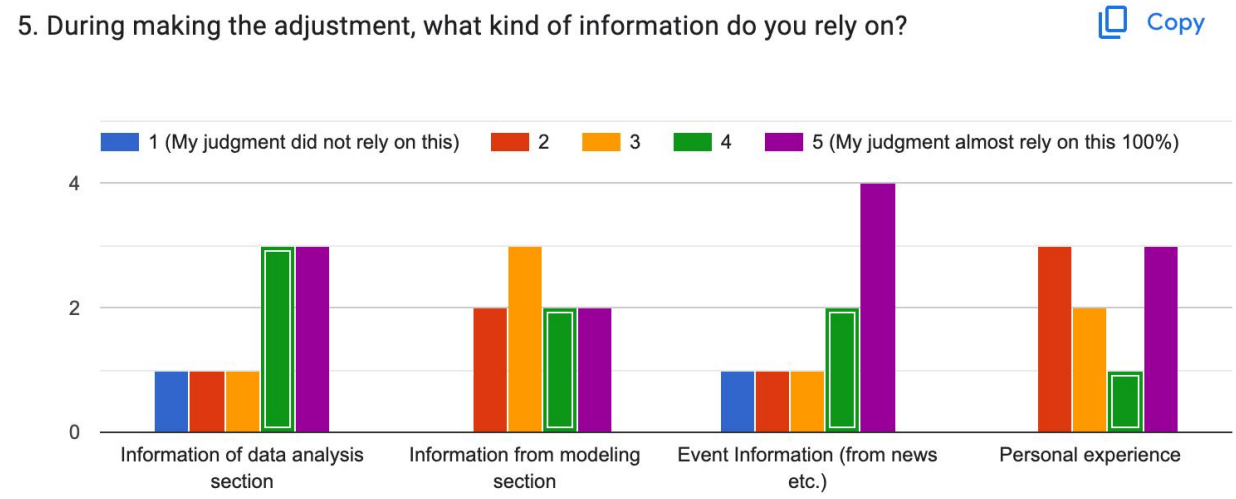

Figure 8.18 Appendix A. Survey results: what kind of information do you rely on

Acknowledgements

前向きな行動から日々新しいことが生まれます。未来を考えることによって、すべてのことがまだ経験していないものとして扱われるため、常に規格を破ることに勇気を持ちます。

- リアン・チーチョウ

Just as it is said, "Life is a journey of cultivation," I believe my academic career is a journey of cultivation too, and this journey is still ongoing. In this journey, I have introspected many of my shortcomings and perceived the direction of my future life. Through these years of research, I have pursued an interesting and unique life while gradually understanding the world and myself. In this journey, I have found something worth spending more time and energy on, which is my current research. I will not temporarily call this thing that I am willing to put energy into as interest, because interest is 70% persistence, 20% ideal, and 10% feedback from results. I will continue to maintain this interest.

In the journey of my life, I spent 7 years in Australia, and Japan is my second place of cultivation. So first of all, I need to thank Professor Kazuo Hiekata for his encounter, which opened the door to life in Japan. If the source of everything is my desire to go to Japan, then the beginning of everything is this encounter. In these two years, I have learned a lot in Japan, not only about academic research, but also

about unique experiences in life, and found the direction of my future life and the interest I want to stick to. I still need to thank Hiekata sensei. Under his guidance, the value in my life can be reflected, and my future direction can be revealed through layers of fog.

At the same time, I want to thank the secretary of the laboratory, Ms. Kazuko Yamamoto, for her frequent help in handling academic matters and other things that I cannot handle alone. Without her, there would be no smooth academic career.

Next, I want to thank my mother, Jun Wang, and my family. Without her silent support and enlightened views, I would not have the choices I have now. I am a lucky person. I can freely choose what I want to persist in. Having my mother support me behind my back is my greatest pride. Although my mother did not guide me in my choices, she has always silently supported me behind her, which gives me great motivation and energy to complete my journey.

Next, I want to thank the members of my laboratory. Ira san, Nonomura san, Wang san, sugita san, kushibuchi san, Torii san and Taskia san. They not only accompanied my research journey, but also witnessed the progress of my research from scratch. In this process, they gave me a lot of support, including listening to my research at every lab meeting, and comments and feedbacks on my research.

Finally, thank you to my undergraduate classmate, Chu san. Along the way, my life has been greatly enriched because of her. Thank you for her existence!

Of course, I also want to thank myself, for persevering all the way, never giving up in the face of every difficulty.

振り返ると、軽い舟は既に数万重山を越えていました。 - 李白