MDPI

Article

Analyzing Long-Term and High Instantaneous Power Consumption of Buildings from Smart Meter Big Data with Deep Learning and Knowledge Graph Techniques

Ru-Guan Wang [1], Wen-Jen Ho [2], Kuei-Chun Chiang [2], Yung-Chieh Hung [2], Jen-Kuo Tai [1], Jia-Cheng Tan [1], Mei-Ling Chuang [1], Chi-Yun Ke [1], Yi-Fan Chien [1,3], An-Ping Jeng [1,3] and Chien-Cheng Chou [1,*]

[1] Information Technology for Disaster Prevention (IT) Program, Department of Civil Engineering, National Central University, Taoyuan 32001, Taiwan
[2] Digital Transformation, Institute for Information Industry, Taipei 10574, Taiwan
[3] Taoyuan Fire Department, Taoyuan 33054, Taiwan
[*] Correspondence: ccchou@ncu.edu.tw; Tel.: +886-3-422-7151 (ext. 34132)

Abstract: In the context of the growing emphasis on energy conservation and carbon reduction, the widespread deployment of smart meters in residential and commercial buildings is instrumental in promoting electricity savings. In Taiwan, local governments are actively promoting the installation of smart meters, empowering residents to monitor their electricity consumption and detect abnormal usage patterns, thus mitigating the risk of electrical fires. This safety-oriented approach is a significant driver behind the adoption of smart meters. However, the analysis of the substantial data generated by these meters necessitates pre-processing to address anomalies. Presently, these data primarily serve billing calculations or the extraction of power-saving patterns through big data analytics. To address these challenges, this study proposes a comprehensive approach that integrates a relational database for storing electricity consumption data with knowledge graphs. This integrated method effectively addresses data scarcity at various time scales and identifies prolonged periods of excessive electricity consumption, enabling timely alerts to residents for specific appliance shutdowns. Deep learning techniques are employed to analyze historical consumption data and real-time smart meter readings, with the goal of identifying and mitigating hazardous usage behavior, consequently reducing the risk of electrical fires. The research includes numerical values and text-based predictions for a comprehensive evaluation, utilizing data from ten Taiwanese households in 2022. The anticipated outcome is an improvement in household electrical safety and enhanced energy efficiency.

Keywords: smart meter data analytics; temporal database; deep learning

1. Introduction

Smart meters were initially developed in the 1970s as devices that utilize digital signals to measure the amount of household power consumption and transmit the data to dedicated databases or data storage systems [1]. Unlike traditional mechanical meters, smart meters offer several advantages, including higher accuracy (with a margin of error of approximately 0.5%, compared to around 2% for traditional meters) [1], more frequent access to power consumption data (with the ability to provide data at second or minute intervals, whereas traditional meters are typically read once a month or every two months), and data communication capabilities [2]. Indeed, smart meters are part of the broader framework known as the smart grid, which encompasses a wide range of new technologies, such as power transmission, distribution, usage monitoring, and pricing, within the power system [3]. Smart meters play a pivotal role in consumption monitoring, positioning them as essential tools for power usage pricing for the smart grid [4]. Notably, the installation of low-voltage smart meters in residential and commercial buildings enables residents to continuously monitor their power consumption, serving as an initial step toward promoting

global energy conservation and carbon reduction efforts [5]. For instance, the United Kingdom has set a policy that by 2025, smart meters will be the exclusive choice for residential and commercial environments, with traditional meters being phased out [6]. Similarly, other major countries have also formulated deployment plans for smart meters.

While the widespread adoption of smart meters holds great promise, achieving comprehensive energy-saving goals might be hindered by low resident participation [7]. To mitigate this concern, leveraging smart meters to offer residents valuable information becomes crucial [7,8]. Smart meters have the capability to gather detailed electricity consumption data from residents. These data hold the potential to provide valuable insights, including predicting future electricity usage and assessing consumption safety. Such insights could significantly enhance the adoption of smart meters [3]. Nonetheless, ongoing research emphasizes that even sophisticated deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models face limitations and challenges, particularly in achieving accurate forecasts of residents' impending electricity consumption [9–12]. Considering these complexities, this study takes a divergent approach, deviating from the pursuit of precise future consumption predictions. Instead, it focuses on identifying instances of potentially hazardous electricity consumption and aims to provide early alerts when households are on the brink of engaging in unsafe electricity usage. This approach not only ensures the safety of residents' lives and property but also cultivates a stronger acceptance and utilization of smart meters.

Hence, this article presents an in-depth analysis of smart meter big data obtained from both residential and commercial buildings in Taiwan. Specifically, Taiwan's Ministry of the Interior's Fire Department has recorded an average of 2000 fire incidents from 2007 to 2022, with building fires constituting a significant 69% of these cases. Within building fire incidents, electrical fires account for a substantial 33% of the total. These fires often result from prolonged exposure of electrical devices or wires to excessively high currents or sudden surges, triggering ignition of nearby flammable materials and causing severe damage. In this context, the study utilizes up-to-date smart meter datasets and employs a series of data pre-processing and customized deep learning techniques to extract relevant information. These advanced deep learning approaches align with recent trends in the literature focused on the analytics of smart meter big data [9,13]. Therefore, the aim of this study is to forecast instances of abnormally high power consumption within minutes, facilitating prompt notifications to residents for preemptive measures. Furthermore, this objective encompasses the identification of patterns that signify unsafe electricity consumption, leading to early recognition and heightened resident awareness. To this end, the manuscript is structured as follows: Section 2 provides a literature review, followed by an explanation of the data pre-processing steps in Section 3. Section 4 presents the analysis results and the evaluation of the deep learning method. Finally, Section 5 concludes the study and provides suggestions for future research.

2. Related Work

2.1. Challenges in the Smart Meter Big Data Analytics

In reality, a smart meter is typically installed at the main power switch of a household, enabling the measurement and transmission of electricity consumption data at frequent intervals, ranging from every few seconds to minutes [8]. In certain cases, multiple meters may be deployed collectively, with each meter capturing the electricity usage of a circuit or a major electrical appliance, such as a refrigerator, air conditioner, washing machine, and more [14,15]. Undoubtedly, the aggregation of household electricity consumption data has catalyzed the rise of big data analytics in recent years [16,17]. Various institutions, such as the National Science Foundation of the United States, the Engineering and Science Research Institute of the United Kingdom, and the Smart City Innovation Center of Denmark, have supported numerous studies on the smart meter big data analytics [18]. The number of related papers has been steadily increasing since 2012. For example, the following discussion delves into three significant themes drawn from the literature: (1) Load management,

which leverages load forecasting as a foundation. The prediction of household electricity consumption enables the appropriate classification and management of different groups of households from the demand side [1,19]. For instance, during periods of tight power supply, households that are more suitable for reducing electricity consumption can be identified. Alternatively, time-based electricity pricing mechanisms can be employed to incentivize households to save electricity during specific time slots. (2) Load characteristic analysis, which involves the grouping of electricity consumption behavior based on the extensive data derived from household electricity consumption. This analysis facilitates comparative assessments among peers, enabling a deeper understanding of consumption patterns [20,21]. Activities such as morning washing or evening meal preparation can be categorized to identify common trends and patterns. (3) Electricity theft detection, which is highlighted as a distinct aspect of the smart meter big data analytics. This detection mechanism focuses on identifying prolonged instances of illicit electricity consumption, serving as a specialized form of abnormal data detection [22,23]. This analysis requires a more extensive duration of electricity consumption data, which is of particular interest to power companies aiming to mitigate losses from unauthorized consumption.

In addition, it is important to note that different brands of smart meters may exhibit variations in the quality of measurement and/or transmission [24]. Consequently, in the domain of managing and analyzing smart meter big data, various challenges pertaining to data storage and pre-processing warrant discussion. To illustrate this, envision a scenario where a smart meter captures power consumption data every second, resulting in the generation of 31,536,000 records per year. This underscores the significant data volume that smart meters can produce. Therefore, it is imperative to devise effective strategies for data storage, management, data pre-processing, and analysis methods to ensure optimal analysis outcomes for such a substantial volume of data [24].

At present, two primary technologies are utilized for storing and managing such big data [8,19–21,25]. The first is the relational database, which is the most prevalent and well-established technology. SQL serves as the recognized standard data query and processing language interface for this database type. The second technology, exemplified by Apache Hive 3.13, is the distributed database, which is exceptionally suited for managing a substantial and continually expanding volume of data [26]. It leverages Hadoop Distributed File Systems to seamlessly integrate additional databases as the data size increases [26]. Moreover, when considering the perspective of smart meter manufacturers, it is important to note that majority of devices currently generate power consumption data in the CSV format. Consequently, the use of programming scripts is imperative for data pre-processing, converting CSV records into one of the previously mentioned database systems [21]. When deciding between a relational or distributed database for smart meter data, it is essential to recognize that this study focuses on a building with fewer than 100 households, each equipped with a smart meter. Despite the substantial data volume, managing it typically remains feasible without necessitating multiple database servers to handle advanced functions, such as load balancing [8]. Therefore, in such a setting, opting for a relational database such as PostgreSQL, as opposed to a distributed option such as Apache Hive, might be the most optimal choice for storing smart meter data. This preference arises from the relational database's proficiency in handling programming scripts for diverse data pre-processing tasks [19,20,25]. Additionally, the main focus of smart meter big data analytics often revolves around electricity consumption records of individual buildings over a maximum three-year period. Such dataset sizes comfortably fit within the capacities of a relational database, offering cost benefits, ease of deployment, and management efficiency. Another prominent tool in the realm of big data analytics is Apache Spark 3.2.4, an analytics framework built on a database foundation, engineered to expedite data querying and processing [27]. However, Apache Spark demands a substantial memory allocation to function optimally [27]. In the context assumed in this study, where a building might have just one database server for all electricity consumption data, introducing an additional server to accommodate Apache Spark's high memory needs might not be the

most economical approach. In lieu of this, utilizing a relational database with standard SQL and self-developed data-processing programs seems more than adequate for managing various tasks across the outlined scenarios.

Furthermore, the time granularity of electricity consumption records, especially generated by different brands of smart meters, often exhibits significant variation [28,29]. For example, occasional instances of missing records may occur, leading to situations where only one or two power consumption records are available per 15 min, despite the intended frequency of consumption measurement being at the minute level [30]. Consequently, within a specific time interval, such as 15 min, the quantity of accurately recorded electricity usage records can fluctuate [30]. The literature often discusses the sampling frequency of smart meters, which can vary significantly, ranging from as fast as thousands of samples per second (expressed in kHz) to as long as two hours [28,29]. Instant electricity consumption data collected over a few minutes are usually sufficient for most analyses, aiding residents in monitoring and conserving electricity usage. However, higher sampling frequencies are recommended for electricity bill pricing, considering the varying periods and progressive tariff structures [7]. In fact, managing the large storage space and complexities associated with electricity consumption records has been a topic of discussion, with proposed methods for maintaining accuracy and reducing the data volume [8,31]. Therefore, prior to conducting the smart meter big data analytics, it is necessary to pre-process the raw data to ensure the presence of an electricity consumption record on the time axis every 5 or 15 min [32,33]. Moreover, when analyzing the electricity consumption behavior of residents, it may be necessary to prepare a separate dataset comprising electricity consumption records at analysis time intervals of 30 or 60 min [32,33]. Essentially, the time granularity of the data collection should be determined based on the analysis algorithm employed [34]. The original electricity consumption records should be adequately preserved in the database, while the generation of electricity consumption data at different time granularities should be realized in real time through a method akin to a database view.

In summary, analyzing the electricity consumption records for all buildings together at the city level may raise concerns regarding the potential breach of personal privacy [25]. Further, for electricity consumption records of five years or longer, it seems more suitable for the power company to store and analyze such huge datasets. For the database server at the building or household level, it may be better to manage the electricity consumption records within a three-year timeframe. Hence, it is recommended to deploy a distributed database on the power company's side, while the server at the building or household level, where the household is located, can utilize a contemporary relational database. This configuration is deemed sufficient for handling the voluminous smart-meter-generated big data within the given assumptions.

2.2. Difficulties of Predicting Household-Level Power Consumption

It is important to recognize that the mere installation of smart meters does not directly lead to reduced electricity bills for residents [24]. To achieve significant energy savings, it is essential to combine smart meters with comprehensive information services that can effectively influence residents' behavior and promote electricity conservation [35]. One common example of such services is electricity consumption forecasting, combined with time-of-use (TOU) rates. This service can suggest optimal usage times for electrical appliances, encouraging households to adjust their consumption habits to save money [7,24,36,37]. Indeed, electricity consumption prediction has long been a focus of research in the field of energy conservation and carbon reduction [38]. Initially, it emerged from energy consumption simulations during the building design stage, estimating air conditioning or lighting energy usage based on building characteristics such as orientation and the window opening rate [38,39]. With the proliferation of Internet of Things (IoT) sensors, various algorithms have been developed in the literature to estimate the electricity consumption of buildings or households using sensed data [30,40–42]. In addition, the rapid advancements in artificial intelligence technology in recent years have contributed to more accurate household

electricity consumption prediction models, aligning them better with real-world needs. It is through the combination of smart meters, advanced prediction techniques, and behavioral changes that substantial progress can be made in achieving energy savings and promoting sustainable practices. Certainly, load forecasting entails the prediction of future power consumption, and this can be accomplished through two primary methods: (1) predicting household behavior by leveraging various sensors (e.g., motion sensors, and temperature and humidity sensors) to estimate power consumption, sometimes on a per-appliance basis [7,31,36,37], and (2) directly utilizing extensive electricity consumption data from smart meters to project future consumption [19].

Despite the clear demand for load forecasting and the availability of numerous methods [1], achieving a high forecasting accuracy remains a challenge. Some studies have shown forecasting errors reaching up to 300% [8]. In fact, there are several metrics available to assess the accuracy of electricity consumption forecasting, broadly classified into two categories. The first category is suitable for comparing the accuracy of different forecasting methods within the same dataset, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The second category is applicable across various forecasting methods, reflecting data collection and expressing the accuracy of each prediction method as a percentage, for instance, Mean Absolute Percentage Error (MAPE), as shown in Equation (1):

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - y_{pred,t}}{y_t}\right| \tag{1}$$

where t represents a specific record in the electricity consumption dataset, n represents the total number of records, y_t represents the actual power consumption value, and $y_{pred,t}$ represents the predicted power consumption value.

Hence, MAPE stands out as an adequate indicator for comparing error levels among various prediction methods across different electricity consumption datasets. Moreover, considering the wide array of prediction techniques and metrics utilized in the literature, the research team conducted a review of three recent articles that employed deep learning methods for electricity consumption prediction, all of which employed the MAPE metric to demonstrate prediction performance. The first article employs the LSTM method to forecast the power consumption of air-conditioning systems in multiple factory settings, achieving a MAPE value of approximately 10% [10]. The second article utilizes a general deep learning approach as well as a custom CNN and LSTM method for predicting power consumption in the well-known IHEPC public household power usage dataset, yielding a MAPE value of around 30% [11]. The third article applies the LSTM method to predict real-time electricity consumption in school buildings, resulting in accuracy ranging from 5% to 30% on various occasions [12].

The studies reported in [10,12] highlighted that when electricity consumption data are collected in relatively simple environments, such as factory air-conditioning systems or office settings such as schools, the MAPE values are typically small. However, there still exists a 10% MAPE error, primarily attributed to high electricity consumption instances [10,12]. In essence, when focusing solely on the high-power consumption records in the dataset, such as predicting consumption exceeding 3000 W, the MAPE value tends to increase due to the limited data points available. Conversely, when analyzing electricity consumption data from typical households, even with the use of enhanced deep learning techniques, the MAPE value remains high, around 30%. Consequently, it is affirmed in the literature that predicting electricity consumption for general households presents a difficult challenge. To effectively address this, it is imperative to devise robust data pre-processing strategies and formulate predictive algorithms tailored to the specific forecasting needs of households, ensuring that residents can fully leverage the advantages of smart meters.

Finally, the literature shows that when analyzing electricity consumption at the city or community level, the overall prediction accuracy tends to be relatively high due to the larger scale [25]. On the other hand, when the analysis is performed at the household

level, while it provides the most relevant insights for individual households, the prediction accuracy is not always ideal [40]. To improve the accuracy of household-level electricity consumption prediction, it is necessary to augment smart meter data with additional environmental factors (such as temperature and humidity, illuminance, etc.), building characteristics (such as orientation, indoor area, building materials, etc.) [41,43,44], and the information pertaining to residents' daily schedules and activities [45]. However, collecting and analyzing additional personal privacy data to enhance the precision of electricity consumption forecasting may raise concerns and potentially discourage residents from installing smart meters [24,34]. Further, recent literature indicates that, in order to enhance accuracy, it is necessary to integrate a greater number of sensors for monitoring buildings, residential environments, and residents' behavior [37]. Nevertheless, this raises concerns about personal privacy, making it difficult for the public to accept such intrusive monitoring practices [34].

2.3. The Need of Predicting Unsafe Power Usage Events

Overall, the existing literature on the big data analytics of smart meters primarily focuses on power consumption prediction. Other applications are relatively scarce, possibly due to the recent implementation of smart meters in developed countries and the gradual accumulation of electricity consumption data [3]. However, British scholars have explored the relationship between smart meters and residential fires, indicating a slight correlation resulting from flaws in the process of replacing old meters with new smart meters, leading to incomplete wiring and subsequent fire incidents [3]. From the perspective of residents' concerns, electrical safety is undoubtedly one of the most important issues. Nevertheless, there is currently limited literature utilizing smart meter big data to predict unsafe electricity consumption that can lead to electrical fires.

Predicting instances of unsafe electricity usage within a household is comparatively less complex than the task of household-level electricity consumption prediction, as it can solely rely on the household's historical smart meter big data. Moreover, such predictive capabilities offer significant advantages to households while requiring a lesser amount of personal privacy information. The underlying assumption for predicting unsafe electricity consumption is similar to load prediction, assuming no significant changes in residents' behavior (e.g., prolonged absences or tenant turnover), enabling algorithms to utilize past electricity consumption data to forecast future patterns. Previous literature has explored abnormal power consumption from two main perspectives: the power company side and the user side [46]. The power company side focuses on identifying cases of electricity theft and detecting discrepancies between actual and measured power consumption [22,46]. The user side examines abnormal power consumption patterns related to appliances and explores strategies for power-saving appliance replacements [47]. However, incidents of extremely high power consumption that lead to actual electrical fires are rare. As a result, it is not feasible to solely focus on expanding the records of extremely high power consumption. Similarly, deleting the records of normal power consumption is also not appropriate.

Thus, it is believed that the integration of such warning mechanisms into smart meters could offer a valuable capability to promptly notify residents about potentially unsafe electricity usage. This would empower residents to take timely corrective actions, thereby ensuring the safety of their homes. Prioritizing these advanced predictive features should be regarded as a fundamental service provided by smart meters [7], ultimately enhancing residents' confidence in their effectiveness. This increased trust is expected to stimulate higher rates of smart meter adoption.

3. Identification of Long-Term Power Consumption Patterns

3.1. Data Format of Smart Meter Data

Although the amount of data generated by a smart meter is usually large, the format of such data is quite simple. Taking the CSV format of smart meters deployed in Taipei City as an example, Table 1 lists the four fields with their data types and descriptions.

Table 1. The table schema for power consumption data generated by smart meters.

Field Name	Data Type	Descriptions
house_id	varchar (50)	Representing household ID number
w	float	Representing the wattage value measured
reporttime	datetime	Representing the report time
channel_id	varchar (10)	Representing the circuit ID number. Note that "0" represents the main switch meter.

In Taiwan, the standard voltage is 110 V, with a frequency of 60 Hz. For this study, the research team selected several buildings in Taipei City and collected power consumption data from 450 households spanning the years 2021 and 2022. Using December 2021 as an example, the electricity consumption records for this period were stored in a CSV file, which had a size of approximately 468 MB. The CSV data were then transferred to a table in a relational database, such as open-source PostgreSQL (see Table 1 for the table's schema). In total, there were 6,910,393 records stored in the database table, including measurements from several indoor circuits, in addition to the main switch. If only the power consumption records of the main switch are considered (represented by channel_id = 0), Table 2 presents some records of one household on 1 December 2021. It can be observed that the time intervals at which the smart meter measures the master switch are not fixed, ranging from approximately 14 s to 130 s between records. Therefore, assuming a fixed time interval for representing electricity consumption records by a smart meter is not reasonable. It is crucial to preserve the original electricity consumption records in the database and generate records with an appropriate time granularity as needed during the analysis process.

Table 2. Sample electricity consumption records of a household on 1 December 2021 in chronological order.

House_id	W	Reporttime	Channel_id
Home01	304.97	2021-12-01 00:00:19.000	0
Home01	304.97	2021-12-01 00:00:34.000	0
Home01	273.44	2021-12-01 00:01:19.000	0
Home01	230.46	2021-12-01 00:01:51.000	0
Home01	305.58	2021-12-01 00:02:25.000	0
Home01	305.58	2021-12-01 00:02:41.000	0
Home01	229.93	2021-12-01 00:02:55.000	0
Home01	229.93	2021-12-01 00:03:11.000	0
Home01	229.93	2021-12-01 00:03:25.000	0
Home01	229.93	2021-12-01 00:03:41.000	0

3.2. Pre-Processing of Time Granularity of Smart Meter Data

Based on the previous review of the literature and the research team's experience in analyzing electricity consumption records, the utilization of big data from smart meters to analyze residents' electricity consumption behavior often encounters the following issues:

- Time interval variations. While the wattage value of each electricity consumption record is typically reliable (with an error rate usually below 0.5% [1]), it is important to note that there can be discrepancies in the time intervals between these records, which can range from a few seconds to minutes.
- Long-term behavior analysis. To accurately analyze long-term behavior, such as air conditioner usage, power consumption records with larger analysis time intervals, such as one record per hour, are needed. Conversely, for activities such as entertainment, power consumption records with shorter analysis time intervals, such as every

15 min, are required. Here, the term time interval is defined in this research as the time difference between two actual power consumption records, while the term analysis time interval refers to the time span covered by the two power consumption records used for analysis (typically derived from multiple actual power consumption records). Hence, due to the disparity between the analysis time interval required for power consumption behavior, which is typically larger (e.g., one record per hour), and the smaller time interval at which electricity consumption records are actually captured (e.g., one record per minute), decision makers must determine the number of actual power consumption records required to establish a representative value for the analysis time interval. For instance, if there are 45 power consumption records between 3 PM and 4 PM with an average value of 150 W, and a threshold of more than 40 power consumption records is set, it can be concluded that the power consumption remained constant at 150 W during that hour. However, if there are only 10 power consumption records within the interval, the decision-maker must provide a conversion formula, such as dividing the average wattage value of the 10 power consumption records by 6, to estimate the electricity consumption value for this analysis time interval. In other words, in cases where the number of electricity consumption records within an analysis time interval is inadequate, decision-makers must offer a conversion formula to estimate the average electricity consumption for that specific analysis time interval.

- Managing missing records. Certain analysis time intervals may lack any electricity consumption records [48]. During analysis, it is necessary to account for these intervals, displaying zero as the electricity consumption value.

The data pre-processing method proposed in this study is described as follows: To begin, the decision-maker analyzing the big data from smart meters needs to determine the analysis time interval, such as requiring an aggregated electricity consumption record every 15 min. The decision-maker then establishes these time points at intervals equal to the analysis time interval, based on the start and end dates of the analysis, as shown in the database table, dt. For instance, assume that the analysis starts at 0:0:0 on 1 October 2021, and ends at 0:0:0 on 1 November 2021. In dt, the second column of each record represents the time value, with a difference of 15 min, while the first column of each record is the serial number, which is recommended to start from 1, with an increment of 1 for each record.

Subsequently, the decision-maker determines the required number of power consumption records (referred to as a) within the analysis time interval. This enables the average wattage of these records to be considered as the representative wattage value for the analysis time interval. In cases where the number of electricity consumption records is below a during the analysis time interval, the average wattage of these records is adjusted by multiplying it with a reduction factor b to represent the representative wattage value within the analysis time interval. Algorithm 1 illustrates the SQL command for creating a database view that lists all power consumption records between the specified analysis start and end dates, adhering to the defined analysis time interval.

In essence, Algorithm 1 is divided into three separate SQL statements that are executed individually, and their results are combined (via SQL union) to create the final view. The first sub-SQL (Line 2) categorizes the original power consumption records based on the analysis time interval specified in the dt table. If the number of records within the interval is equal to or greater than a, the average value is directly considered as the wattage value. The second sub-SQL (Line 4) operates in a similar manner to the first sub-SQL, but for intervals with a smaller number of records, the coefficient b is multiplied. The third sub-SQL (Lines 6–7) consists of two instructions, where Line 7 identifies all the analysis time intervals that contain corresponding power consumption records, and Line 6 performs the set difference operator to identify the analysis time intervals without any power consumption records.

Once the original power consumption records have been pre-processed using Algorithm 1, it can be observed that each analysis time interval is assigned a wattage value, making it more suitable for subsequent analyses without concerns about interval interruptions or data gaps. Furthermore, the inclusion of various SQL 'where' commands

in Algorithm 1 ensures that the wattage value must be greater than or equal to zero, thereby mitigating the risk of incorporating highly biased power consumption records into subsequent analysis considerations.

Algorithm 1. Creating a database view to pre-process all power consumption records between the specified analysis start and end dates for a given household.

Input:

> *dt*: The database table containing all the time points between the analysis start and end dates, with each interval equal to the given analysis time interval.
> *a*: The minimum number of power consumption records within the analysis time interval, necessary for utilizing their average as the representative wattage value.
> *b*: The reduction factor to be applied to the average wattage of the power consumption records within the analysis time interval, in cases where the number of records is less than *a*.
> *houseID*: The identification of the household.
> *tblName*: The name of the database table containing the actual power consumption records, typically imported from a CSV file.

Output:

> A database view consisting of three columns: (1) the first column represents the time point, indicating the start time of each analysis time interval, (2) the second column denotes the number of actual power consumption records within each analysis time interval, and (3) the third column contains the representative wattage value for each analysis time interval.

SQL:

1. create view *houseID*_view as (
2. select d1.t as t, COUNT(*) as iCount, AVG(o.w) as w from *dt* as d1, *dt* as d2, *tblName* as o where o.house_id = '*houseID*' and d1.id + 1 = d2.id and o.w > = 0 and o.reporttime > = d1.t and o.reporttime < d2.t and o.channel_id = 0 group by d1.t, o.channel_id having count(*) > = *a*
3. union
4. select d1.t as t, COUNT(*) as iCount, AVG(o.w)*b* as w from *dt* as d1, *dt* as d2, *tblName* as o where o.house_id = '*houseID*' and d1.id + 1 = d2.id and o.w > = 0 and o.reporttime > = d1.t and o.reporttime < d2.t and o.channel_id = 0 group by d1.t, o.channel_id having count(*) < *a*
5. union
6. select d1.t as t, 0 as iCount, 0.0 as w from *dt* as d1, *dt* as d2 where d1.id + 1 = d2.id and d1.t not in (
7. select d1.t as t from *dt* as d1, *dt* as d2, *tblName* as o where o.house_id = '*houseID*' and d1.id + 1 = d2.id and o.reporttime > = d1.t and o.reporttime < d2.t and o.channel_id = 0 group by d1.t, o.channel_id
8.))

3.3. Temporal Coalescing: Long-Term Power Consumption Patterns

In the realm of smart meter big data, instances of extremely high power consumption (e.g., 4000 W) within a specific analysis time interval can pose a significant risk of electrical fires, particularly when preceding periods also exhibit somewhat high consumption (e.g., 2000 W) patterns. The following section will delve into the prediction of extremely high power consumption, but before that, it is imperative to identify such prolonged periods characterized by long-term high power consumption from the vast power consumption dataset. The advantage of identifying these prolonged periods of high power consumption lies not only in alerting residents to prioritize electrical safety but also in utilizing these periods, along with subsequent periods (whether exhibiting extremely high power consumption or not), as training or testing datasets for the deep learning model described in Section 4. Such identification could help enhance the prediction model's accuracy.

In this study, a prolonged period is defined as a duration lasting for a minimum of two minutes or a specified number of minutes. During this period, there must be at least one actual electricity consumption record captured within each minute, and its wattage

value surpasses a predetermined threshold, such as 2000 W. In other words, within the same minute, there may be multiple electricity consumption records, and at least one record must exceed the threshold value, while other records may or may not exceed the threshold. Subsequently, if these one-minute intervals are adjacent to each other, they will be combined to form longer time intervals. Such database operations can be referred to as temporal coalescing, which is the process of merging multiple rows with equivalent values into a single row when their validity periods overlap. Consequently, for these periods lasting two minutes or more, each minute along the time axis contains at least one record of electricity consumption surpassing the threshold value. This indicates a prolonged period of high electricity consumption in households, which is significant and should be noted. Algorithm 2 describes the steps to identify all prolonged periods from the original power consumption records, and Figure 1 shows the sample output of such prolonged periods, based on the actual power consumption records of a household in October 2021.

Algorithm 2. Identifying all prolonged periods with high wattage values in the power consumption records.

Input:

    c1: In the definition of a prolonged period, its length of time must be greater than *c1* min. The default value is 1.

    c2: In the definition of a prolonged period, it is required that there exists a power consumption record for each minute, and the wattage value of that record must exceed a specified threshold value of *c2*.

    houseID: The identification of the household.

    tblName: The name of the database table containing the actual power consumption records, typically imported from a CSV file.

    v3: The name of a temporary database view for the temporal projection operation. *v3* can be deleted after Algorithm 2.

    vv3: The name of a temporary database view for the temporal coalescing operation. *vv3* can be deleted after Algorithm 2.

Output:

    A table (see Figure 1) representing all the prolonged periods and consisting of two columns: (1) the first column, fromT, represents the beginning time point of each prolonged period, and (2) the second column, toT, denotes the end time point of each prolonged period.

SQL:

1.   create view *v3* as select DATETIMEFROMPARTS (DATEPART (yy,t1.reporttime), DATEPART (mm,t1.reporttime), DATEPART (dd,t1.reporttime), DATEPART (hh,t1.reporttime), DATEPART (mi,t1.reporttime), 0,0) as z, count (t1.w) as c from *tblName* as t1 where t1.w > *c2* and t1.channel_id = 0 and t1.house_id = '*houseID*'

2.   group by DATETIMEFROMPARTS (DATEPART (yy,t1.reporttime), DATEPART (mm,t1.reporttime), DATEPART (dd,t1.reporttime), DATEPART (hh,t1.reporttime), DATEPART (mi,t1.reporttime), 0,0)

3.   create view *vv3* as select v11.z as fromT, v12.z as toT from *v3* v11, *v3* v12 where DATEDIFF (mi, v11.z, v12.z) = 1

4.   select F.fromT, L.toT from *vv3* F, *vv3* L where DATEDIFF (mi,F.fromT,L.toT) > *c1* and F.fromT < L.toT and not exists (select * from *vv3* M where F.fromT < M.fromT and M.fromT < = L.toT and not exists

5.   (select * from *vv3* M1 where M1.fromT < M.fromT and M.fromT < = M1.toT)) and not exists (select * from *vv3* M2 where ((M2.fromT < F.fromT and F.fromT < = M2.toT) or (M2.fromT < = L.toT and L.toT < M2.toT))) order by F.fromT

|	fromT	toT
1	2021-10-01 22:02:00.000	2021-10-01 22:04:00.000
2	2021-10-01 23:28:00.000	2021-10-01 23:31:00.000
3	2021-10-03 23:32:00.000	2021-10-03 23:38:00.000
4	2021-10-04 21:46:00.000	2021-10-04 21:48:00.000
5	2021-10-04 21:50:00.000	2021-10-04 21:53:00.000
6	2021-10-05 23:25:00.000	2021-10-05 23:31:00.000
7	2021-10-06 21:35:00.000	2021-10-06 21:37:00.000
8	2021-10-06 22:04:00.000	2021-10-06 22:11:00.000
9	2021-10-07 22:12:00.000	2021-10-07 22:15:00.000
10	2021-10-08 23:38:00.000	2021-10-08 23:44:00.000
11	2021-10-10 00:30:00.000	2021-10-10 00:36:00.000
12	2021-10-10 21:35:00.000	2021-10-10 21:38:00.000

Figure 1. Some sample output prolonged periods created by Algorithm 2.

In Algorithm 2, the primary purpose is to pre-process the actual power consumption records. In Line 1, the seconds part of the records is removed, while considering criteria such as household ID and exceeding the threshold for high power consumption wattage. Line 3 creates a database view that presents all 1 min intervals. Finally, Lines 4–5 execute the temporal coalescing operation, where adjacent records are merged into longer intervals for further analysis [49].

Figure 2 provides a visual representation of time intervals, analysis time intervals, and prolonged periods. Each time interval is determined solely by the electricity consumption records and can vary significantly in duration. In contrast, the analysis time interval is defined based on specific analysis requirements and maintains a fixed duration throughout the analysis process. However, decision-makers have the flexibility to adjust the length of the analysis time interval to generate new datasets for subsequent analysis, such as for sensitivity analysis or trial and error. The prolonged period, derived from the original electricity consumption data, allows decision-makers to set a threshold for high power consumption and determine the minimum duration necessary to qualify as a long-term time segment for such high power consumption events.

In summary, in the analysis of smart meter big data, different time granularities are often employed. For instance, when predicting abnormal electricity consumption, it is desirable to have forecasts at intervals of 15 min or less. Another example is examining the correlation between electricity consumption and outdoor temperature, which requires hourly data aggregation. While the raw data from smart meters must be stored for electricity billing purposes [24], storing the averaged or aggregated consumption records at various time granularities for varied analysis would consume significant disk space. Hence, the recommended approach is real-time calculation, dynamically generated through techniques such as using a database view, as shown in Algorithms 1 and 2. Although the calculation

formula for such analysis is not complex, parameter adjustments are necessary based on the actual analysis requirements.
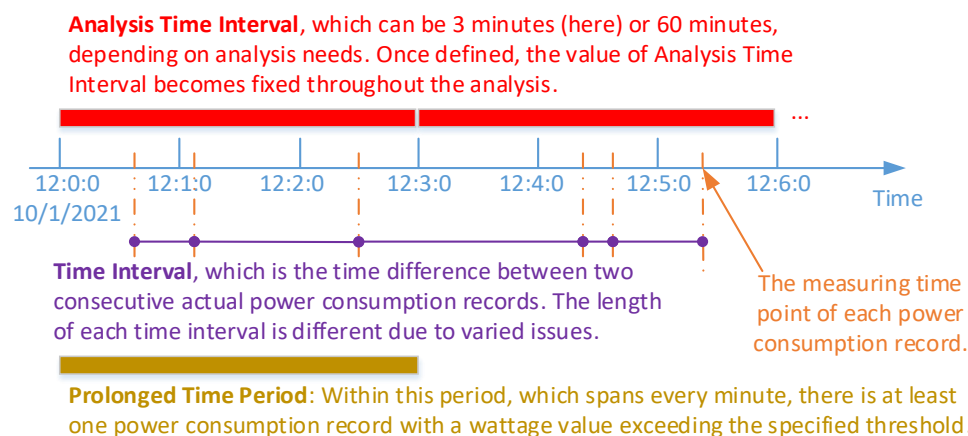
Figure 2. Graphical explanations of time interval, analysis time interval, and prolonged period.

4. Prediction of Extremely High Instantaneous Power Consumption Using Deep Learning

4.1. Data Transformation

Previous research has indicated the challenges associated with predicting residents' electricity consumption behavior. Striking a balance between accuracy and privacy preservation is crucial since excessive collection of personal information may deter residents from adopting smart meters. Nonetheless, from a home safety perspective, leveraging the big data of electricity consumption from smart meters to forecast the likelihood of upcoming instances of extremely high power consumption could prevent electrical fires and enhance resident safety. By providing timely reminders, residents may have greater confidence in the information services offered by smart meters.

In fact, the prediction of imminent extremely high electricity consumption can be viewed as a binary classification problem, i.e., distinguishing between extremely high electricity consumption and not extremely high electricity consumption. This fundamentally differs from predicting the numerical value of future electricity consumption, the challenge that has been discussed in the literature review section. Furthermore, similar to electricity theft analysis, predicting instances of extremely high electricity consumption is not a common occurrence. Therefore, it requires a substantial accumulation of electricity consumption records spanning longer analysis time intervals preceding these historical events to effectively forecast the likelihood of imminent occurrences of extremely high power usage. To address this, the current study proposes a novel approach to transforming numerical electricity consumption records into text format. By employing deep learning algorithms such as RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), or GRU (Gated Recurrent Unit), the study aims to predict the occurrence of future extremely high power usage events based on the textual representation of past electricity consumption records.

The research team initiated the analysis of the acquired electricity consumption data, which encompassed the records from 450 households residing in various buildings within Taipei City during the period of 2021–2022. Among these records, the highest recorded wattage was 21,845.03 W in August 2022. Subsequently, an analysis time interval of five minutes for such predictions was defined, and a random selection of ten households was performed. These households demonstrated a consistent population count between 2021 and 2022, and their smart meter records exhibited no long-term anomalies. For the training process, the research team chose one household's power consumption data from a three-month period of 2021, while the data of the corresponding months in 2022 were utilized as the test dataset. Using the conversion formula specified in Table 3, the electricity consumption records underwent the following procedures: (1) calculation of the average electricity consumption value and the maximum value within each analysis time interval,

and (2) conversion of the average and maximum values into two-letter codes followed by a space, such as 'AA'.

Table 3. The power consumption records conversion table: mapping numerical values of power consumption records to corresponding characters.

ASCII#	Character	Value Range (W)	ASCII#	Character	Value Range (W)
65	A	$0 \leq v < 500$	33	!	$13{,}000 \leq v < 13{,}500$
66	B	$500 \leq v < 1000$	34	"	$13{,}500 \leq v < 14{,}000$
67	C	$1000 \leq v < 1500$	35	#	$14{,}000 \leq v < 14{,}500$
68	D	$1500 \leq v < 2000$	36	$	$14{,}500 \leq v < 15{,}000$
69	E	$2000 \leq v < 2500$	37	%	$15{,}000 \leq v < 15{,}500$
70	F	$2500 \leq v < 3000$	38	&	$15{,}500 \leq v < 16{,}000$
71	G	$3000 \leq v < 3500$	39	'	$16{,}000 \leq v < 16{,}500$
72	H	$3500 \leq v < 4000$	40	(	$16{,}500 \leq v < 17{,}000$
73	I	$4000 \leq v < 4500$	41	)	$17{,}000 \leq v < 17{,}500$
74	J	$4500 \leq v < 5000$	42	*	$17{,}500 \leq v < 18{,}000$
75	K	$5000 \leq v < 5500$	43	+	$18{,}000 \leq v < 18{,}500$
76	L	$5500 \leq v < 6000$	44	,	$18{,}500 \leq v < 19{,}000$
77	M	$6000 \leq v < 6500$	45	-	$19{,}000 \leq v < 19{,}500$
78	N	$6500 \leq v < 7000$	46	.	$19{,}500 \leq v < 20{,}000$
79	O	$7000 \leq v < 7500$	47	/	$20{,}000 \leq v < 20{,}500$
80	P	$7500 \leq v < 8000$	48	0	$20{,}500 \leq v < 21{,}000$
81	Q	$8000 \leq v < 8500$	49	1	$21{,}000 \leq v < 21{,}500$
82	R	$8500 \leq v < 9000$	50	2	$21{,}500 \leq v < 22{,}000$
83	S	$9000 \leq v < 9500$	51	3	$22{,}000 \leq v < 22{,}500$
84	T	$9500 \leq v < 10{,}000$	52	4	$22{,}500 \leq v < 23{,}000$
85	U	$10{,}000 \leq v < 10{,}500$	53	5	$23{,}000 \leq v < 23{,}500$
86	V	$10{,}500 \leq v < 11{,}000$	54	6	$23{,}500 \leq v < 24{,}000$
87	W	$11{,}000 \leq v < 11{,}500$	55	7	$24{,}000 \leq v < 24{,}500$
88	X	$11{,}500 \leq v < 12{,}000$	56	8	$24{,}500 \leq v < 25{,}000$
89	Y	$12{,}000 \leq v < 12{,}500$	57	9	$25{,}000 \leq v < 25{,}500$
90	Z	$12{,}500 \leq v < 13{,}000$	58	:	$25{,}500 \leq v < 26{,}000$

For instance, during the period from October to December 2021, a total of 26,496 analysis time intervals (every five minutes) were created. Considering that each analysis time interval results in two characters and one space character during the transformation process, the resulting text file size for the converted electricity consumption records amounts to 79,488 bytes. The actual electricity consumption records for the 450 households over the 3 months range from approximately 120,000 to 600,000. However, the conversion process is remarkably efficient, with the conversion of each household's records being completed within 20 to 50 s. Figure 3 shows a sample output text file after the conversion process.
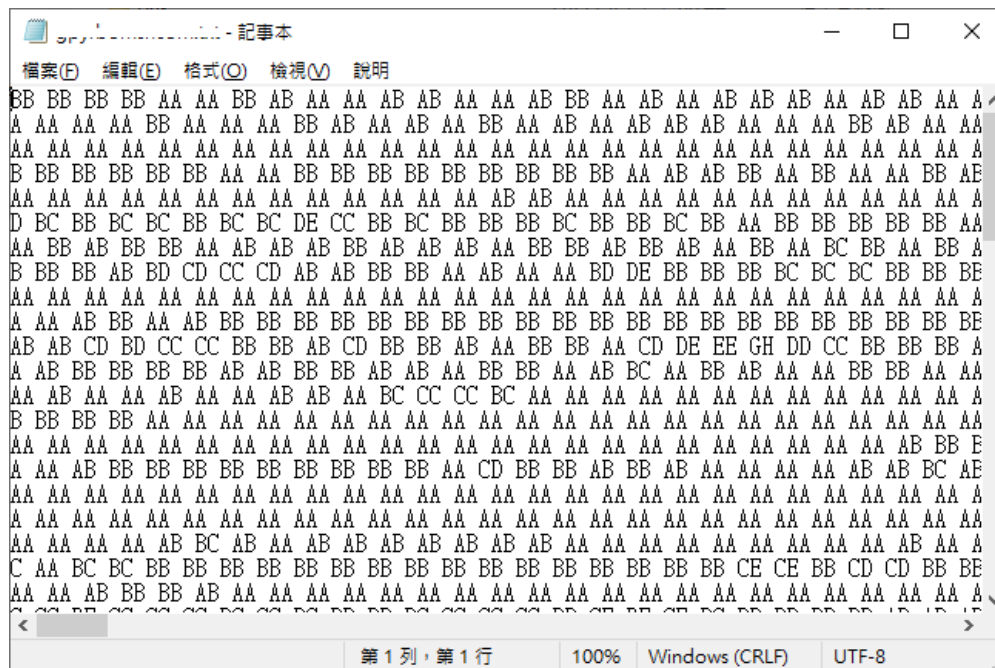
Figure 3. The textual representation of the electricity consumption records for a household from October to December 2021, considering an analysis time interval of five minutes.

4.2. Deep Learning for Predicting Extremely High Power Consumption

This study tackles the prediction problem of identifying potential occurrences of extremely high power consumption events in the near future by employing a two-stage approach. The first stage employs deep learning techniques for text prediction, where the converted text from electricity consumption records serves as input, in order to predict subsequent text. In the second stage, binary classification is performed as post-processing for the first stage. This involves examining the output text from the previous prediction and determining if any characters exceed the threshold for extremely high power consumption. Although the threshold value for extremely high power consumption can be adjusted, it is typically set to a value above 3500 W for a small or medium-sized household. In the context of the converted characters, this corresponds to characters after 'H'. Thus, if any such characters are detected, the result is true; otherwise, it is false.

The deep learning program used in this research is based on Python 3.10.2 with TensorFlow 2.10, utilizing an RTX-3070 GPU for efficient performance. The complete source codes can be provided upon request. It should be noted that the analysis time interval for this prediction is assumed to be five minutes. If decision-makers wish to modify the analysis time interval, they will need to begin with the data pre-processing and transformation steps outlined in the previous section. Essentially, the prediction will be in string format, and the length of the input string, which is customizable by users, will be three times the length of the output string. The default length for the input string is set to 99, which corresponds to 165 min ($99/3 \times 5$).

The prediction program follows the typical workflow of deep learning. Firstly, the data pre-processing codes for transforming numeric power consumption values into the text format are performed. Each analysis time interval is transformed into three characters: an average character, a highest-value character, and a space character. Considering the large amount of data in the input file (a total of 79,488 characters), TensorFlow processes the characters in multiple steps. In the prediction program, with the 'BATCH_SIZE' of 64, each step handles $64 \times (99 + 1)$ characters, resulting in 12 steps ($79,488/64/100$) to finish the analysis. Any remainder of the characters is disregarded.

Figure 4 shows the proposed architecture of the deep learning model. In this model, the first layer is 'Embedding', which encodes each string using the 'embedding_dim' parameter

set to a value of 256. The second layer is the 'Bidirectional' layer, internally utilizing 'GRU' with 1024 units to capture both the forward and backward context of the string, which may have mutual influence. The third layer employs 'GRU' with 256 units to determine the impact of previous strings on subsequent strings. The fourth layer gradually narrows down the model's prediction range, while the fifth layer ensures that the prediction results remain within 'vocab_size'. It should be noted that the 'embedding_dim' parameter represents the number of independent variables used for prediction. The 'EPOCHS' parameter indicates the number of times the deep learning method will iterate over the dataset during training. Typically, increasing the number of epochs can improve the prediction performance up to a certain limit, after which the benefits may diminish. The 'Dropout' parameter, set to a value of 0.2 in this case, can be used to release some variables to prevent overfitting the training dataset. Finally, the 'temperature' parameter, commonly utilized in LSTM-like algorithms, is used to regulate the level of randomness in the predictions. This parameter plays a crucial role in adjusting the balance between high- and low-probability terms within the output distribution. Higher temperature values introduce more randomness and diversity in predictions, whereas lower temperature values enhance predictability by emphasizing high-probability terms and suppressing the influence of low-probability ones.

```
Epoch 25/25
12/12 [==============================] - 1s 116ms/step - loss: 0.0026
Model: "sequential_1"

Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (1, None, 256)            2816

bidirectional_1 (Bidirectio  (1, None, 2048)           7876608
nal)

gru_3 (GRU)                  (1, None, 256)            1771008

dense_2 (Dense)              (1, None, 22)             5654

dropout_1 (Dropout)          (1, None, 22)             0

dense_3 (Dense)              (1, None, 11)             253

=================================================================
Total params: 9,656,339
Trainable params: 9,656,339
```

Figure 4. The summary of the deep learning model with six layers.

The deep learning model depicted in Figure 4 achieves a training result with a loss value of 0.0026 and a total of 9,656,339 variables, representing the electricity consumption of a specific household during the period from October to December 2021. The model utilizes only 11 distinct characters from Table 3. Essentially, this deep learning model predicts the next character based on the input string, with a total of 11 possible choices. This process is repeated multiple times to generate a complete output string.

It is worth noting that when analyzing different months or households, the electricity consumption records can vary significantly, leading to potential differences in the characters used (from Table 3). In this study, a three-month dataset of electricity consumption records was selected as the training data. This duration provides a substantial volume of smart meter data that correspond to a specific season. For application to other households or months, decision-makers have the flexibility to adjust various parameters, such as the analysis time interval, 'analysisTimeIntervalSize' value, and other common parameters in deep learning. Furthermore, with the current combination of datasets and parameters, TensorFlow requires approximately three minutes to complete the training of this model.

If decision-makers are willing to invest more analysis time, they can include additional months to analyze a larger set of electricity consumption records. By doing so, the deep learning method will be able to uncover more intricate power consumption patterns. The extended analysis period allows for the discovery of power consumption patterns that may occur infrequently in the original data, enabling residents to receive early warnings and adopt safer electricity consumption practices. As previously mentioned, not every household has an equal number of unsafe electricity usage records within the designated interval. To address this, the research team employed the generated deep learning model to randomly populate records, ensuring each household had 10 instances of unsafe electricity usage events. These events, along with the electricity consumption records from the preceding 165 min, collectively constitute all the attributes of a node in this configuration.

Finally, the research team leveraged three-month electricity consumption data from five households to construct their individual deep learning models and knowledge graphs. Figure 5 was constructed using Neo4j's knowledge graph tool and portrays the five households (depicted as pink nodes) within a community (indicated by the green node). For the three households (Home02/Home03/Home04) experiencing incidents of unsafe electricity usage, their electricity consumption data are denoted by brown nodes. Their maximum power consumption surpassed the 4000 W threshold, represented by the blue node. The remaining two households did not record such exceedingly high power consumption during the specified three-month period. It should be noted that not every household had an equal number of unsafe electricity usage records within the designated period. To address this imbalance, the research team employed the generated deep learning model to randomly generate records, ensuring each household had 10 instances of unsafe electricity usage events. These events, combined with the electricity consumption records from the preceding 165 min, collectively form the attributes of one node within this knowledge graph configuration.
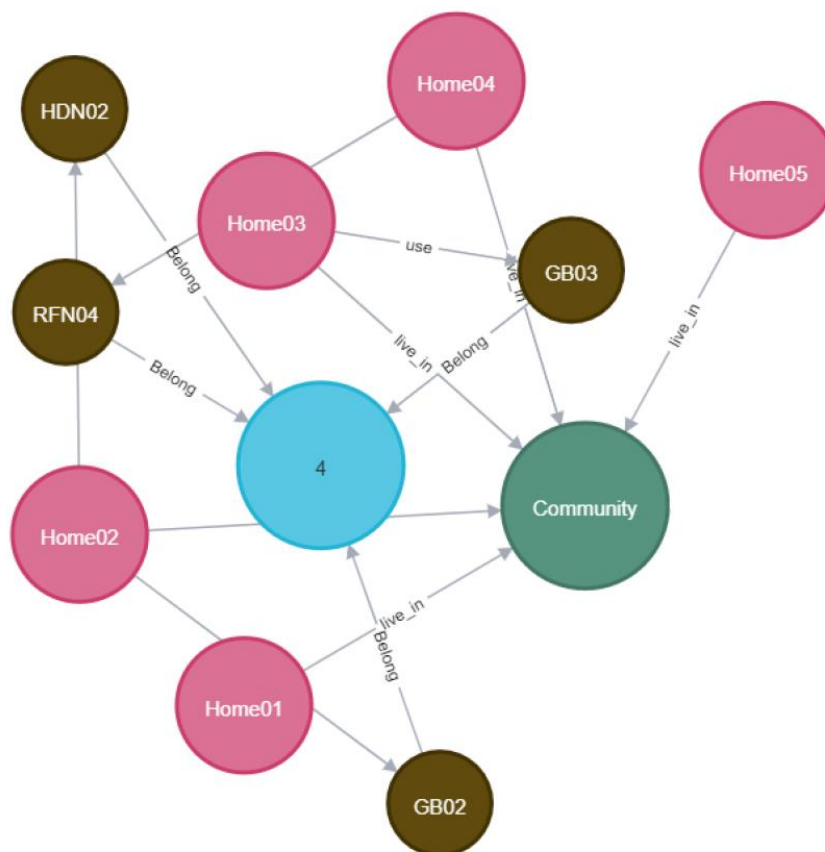
Figure 5. A knowledge graph illustrating multiple households with their power consumption records associated with a 4000 W threshold.

Furthermore, based on the knowledge graph and the tool, the research team utilized the Fast Random Projection (FastRP) algorithm to reveal nearby nodes that share similar electricity consumption patterns, as demonstrated in Figure 6 [50–52]. FastRP allows for aggressive dimensionality reduction while preserving most of the distance information. FastRP operates on graphs, which involve nodes, edges, and attributes, and tries to preserve similarity between nodes and their neighbors. This means that two nodes that have similar neighborhoods should be assigned similar embedding vectors. Conversely, two nodes that are not similar should not be assigned similar embedding vectors. Despite the small number of nodes, FastRP identifies nodes with similar attributes and represents their positions in a two-dimensional plane with x/y coordinates ranging from −1 to +1. In Figure 6, Home03 and Home04 are positioned closest to each other because their attribute values, which represent extremely high power usage, were similar. Consequently, the residents of Home03 and Home04 can be organized to participate in the same electricity safety education and training, or their unsafe electricity usage warning conditions can be collectively monitored. In the event of a future warning condition triggering within the specified range, the residents should be able to receive notifications accordingly.
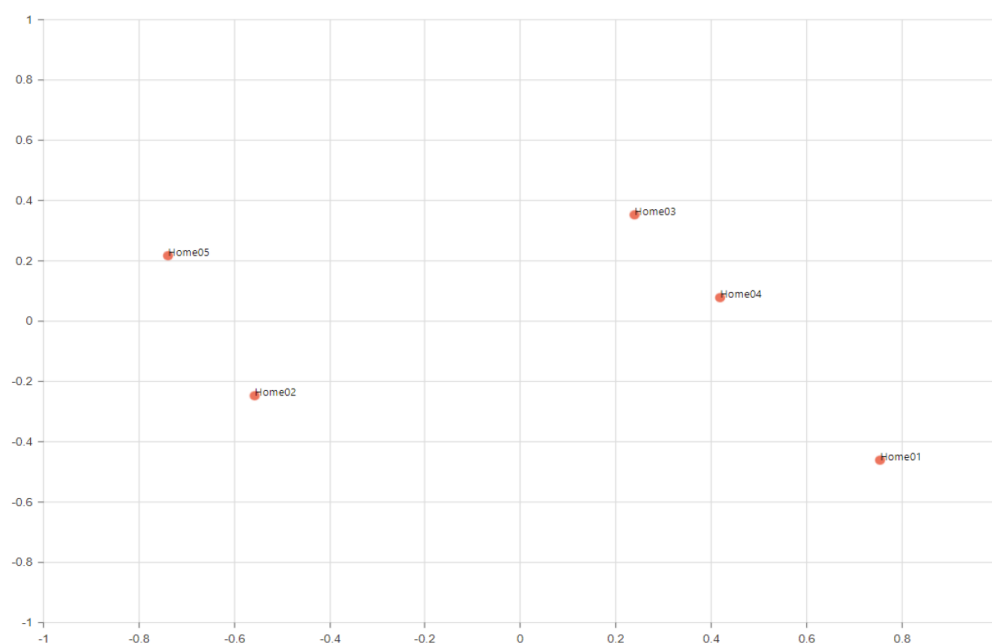
Figure 6. Utilization of FastRP for the identification of households with similar unsafe power usage patterns (Home03 and Home04).

4.3. Prediction Results and Evaluations

The previous section described the utilization of the deep learning method to predict the converted strings of electricity consumption records for the subsequent 55 min (33 characters), based on the converted strings of the current electricity consumption records for the previous 165 min (99 characters). In fact, this model was built upon the electricity consumption data of a certain household in Taipei City during the period of October–December 2021. For model testing, as there were no changes in the selected household's family members or daily schedule from 2021 to 2022, firstly, the research team obtained the CSV file containing the original smart meter electricity consumption records for this household from October to December 2022. Subsequently, the CSV file was imported into the PostgreSQL database, and Algorithm 1 was executed with the analysis time interval set to 5 min, the input parameter *a* set to 2, and the input parameter *b* set to 0.2. The pre-processed data can be accessed at any time through the database view defined in Algorithm 1. Then, referring to Table 3, the database data were exported into a text file in the character format. Within this file, the

research team randomly selected two distinct strings, each representing 220 min (165 + 55) of electricity consumption records during the period of October to December 2022.

It is important to highlight that both such strings had the same length of 132 characters and that within their first 99 characters, there was no occurrence of the character 'H' or any subsequent characters. This indicates that the household did not experience any instances of extremely high electricity consumption in the 165 min leading up to the given time point. However, it is noteworthy that the first string exhibited extremely high power consumption within 55 min following the given time point, whereas the second string did not exhibit any extremely high power consumption events within 55 min following the given time point.

Figure 7 illustrates the input and output of the deep learning prediction program, where the first input string commences with 'AA' and concludes with 'CE EE EE'. In the original electricity consumption records, an immediate occurrence of extremely high power consumption was recorded, indicated by the presence of the character 'H' in the subsequent position. Following the prediction by the program, the generated output string is displayed in the second line of Figure 7. The first 99 characters of the output string match the input string precisely, while the remaining 33 characters are entirely generated by the program. This sequence represents the predicted future electricity consumption. Taking the second line of Figure 7 as an example, the 33 characters are 'HH B AB B GH B AD D FI BF F F CF', encompassing both the 'H' character and 'I' characters. This indicates that the program successfully predicts future instances of extremely high power consumption.

Figure 7. Two input strings and the corresponding output strings: the 'AA'-prefixed string indicates extremely high power consumption, while the 'BB'-prefixed string leads to not extremely high power consumption.

Moving to the third line in Figure 7, it represents an input string that starts with 'BB' and ends with 'BB AB AA'. In the original electricity consumption record, no instances of extremely high power consumption were expected within the subsequent 55 min. After the input string is subjected to prediction by the program, an output string is generated, as depicted in the fourth line of Figure 7. The initial 99 characters of the output string match the input string, and the subsequent 33 characters are completely generated by the program. This output can be regarded as the predicted outcome for the future electricity consumption scenario. Taking the fourth line of Figure 7 as an example, the 33 characters are 'A BB B AB B B BD B B B A A A A A ', and no character 'H' or any subsequent characters are present, indicating the absence of extremely high power consumption.

The research team selected another 10 households from the smart meter big data and analyzed their electricity consumption patterns, and Table 4 provides an overview of these households. In the second column, the power consumption data of the specified months in 2021 were identified as the training dataset, while the corresponding months in 2022 served as the test dataset. The third column represents the total number of original electricity consumption records during the specified three-month period in 2022. During the three-month period, assuming one electricity consumption record per minute, there would theoretically be 129,600 records. However, as shown in Table 4, the smart meter records of these 10 households indicate an average time interval of approximately 4 to 5 records per minute.

In the fourth column, the number of strings extracted from each household's electricity consumption records of extremely high wattage values is indicated. As previously explained, the first 99 characters of each string do not exhibit extremely high power consumption, while the rest of the characters do include extremely high power consumption. Identifying a period without extremely high power consumption was relatively straightforward. However, identifying a period with extremely high power consumption, which may exhibit sustained rather than short occurrences, posed a greater challenge. To address

this, the research team employed Algorithm 2 in order to identify the periods that meet the specified criteria. Subsequently, the relevant text strings of electricity consumption records were extracted for further analysis and prediction. Lastly, the research team utilized the deep learning prediction program to classify each text string, and the fifth column shows the accuracy of each prediction for the 10 households. These assessments reveal that the model's predictive capability exceeds 90%. In practical terms, it can provide residents with warnings up to 165 min before the occurrence of unsafe electricity usage incidents, enabling them to take preventive measures and avoid such incidents.

Table 4. The evaluation results of the 10 households spanning the years 2021 and 2022 using the proposed two-stage approach.

House ID	Months	# of Original Records in 2022	# of High-Wattage Records for Testing	Accuracy
Home01	1–3	519,041	28	0.964
Home02	7–9	519,847	30	0.967
Home03	10–12	518,694	32	0.938
Home04	1–3	508,964	34	0.941
Home05	4–6	508,837	28	0.964
Home06	4–6	508,882	30	0.967
Home07	7–9	508,710	36	0.944
Home08	10–12	508,403	34	0.941
Home09	1–3	508,374	36	0.972
Home10	4–6	508,347	30	0.967

The research team also employed common Conv1D and LSTM methods to directly predict the numeric values of electricity consumption records. For Conv1D, the filter parameter is configured at 256, and the 'kernel_size' parameter is set to 34, corresponding to a time span of (165 + 5) minutes divided into 5 min intervals. The activation function is specified as 'relu', and the final two layers are defined as 'Dense'. In the case of LSTM, the unit parameter is set to 256, and the 'return_sequences' parameter is set to 'True.' The activation function used is 'relu,' and the final layer is 'Dense'. This prediction was performed using the same dataset of electricity consumption from the 10 households. It includes records from the preceding 165 min as input and aims to iteratively predict electricity consumption for the subsequent 5 min, resulting in a total of 55 min of output. In the quantitative analysis, the predicted results were directly compared with actual electricity consumption records, and the highest predicted consumption value was checked for indications of unsafe electricity usage events. The results are summarized in Table 5. Column 2 presents the prediction results for 30 randomly selected electricity consumption records from each household's dataset using the Conv1D method. The MAPE index was used to assess the prediction performance, facilitating comparisons with methods from previous literature. Column 3 utilized the same approach but employed the LSTM method for prediction. From the data presented in Columns 2 and 3 of Table 5, it is evident that the MAPE values for the numerical predictions made using Conv1D and LSTM methods fall within the range of 10% to 15%, which is consistent with findings in the existing literature. However, the challenge arises from the limited number of extremely high electricity consumption records in each household's dataset, making it impractical to predict occurrences of extremely high electricity consumption.

As shown in Columns 4 to 7 of Table 5, when specifically analyzing the records of extremely high electricity consumption (approximately 30 records) from these households, the prediction outcomes significantly differ. Column 4 reveals that using the Conv1D method to predict extremely high power consumption resulted in a notably high MAPE of

around 70%. Employing binary classification directly to determine whether it constitutes an unsafe event yielded an error rate of approximately 80%. Similarly, when LSTM was utilized for prediction, the MAPE remained high at around 60%, and the binary classification error rate was approximately 70%. In essence, relying on the direct predictive values from such methods is inadequate for providing early warnings of unsafe electricity events to households. Finally, the MAE and RMSE values of the aforementioned prediction for the extremely high power consumption records are listed in Table 6. It is evident that directly predicting future extremely high electricity consumption values from past records results in significant errors. This underscores the importance of the data pre-processing steps proposed in this study for achieving accurate results.

Table 5. The numerical predictions and binary classification error rates for the 10 households using common Conv1D and LSTM models.

(1) House ID	(2) MAPE of Using Conv1D for 30 Randomly Selected Records	(3) MAPE of Using LSTM for 30 Randomly Selected Records	Using Conv1D for the Extremely High Power Consumption Records		Using LSTM for the Extremely High Power Consumption Records	
			(4) MAPE	(5) Error of Binary Classification	(6) MAPE	(7) Error of Binary Classification
Home01	0.12	0.11	0.78	0.79	0.58	0.79
Home02	0.11	0.09	0.73	0.79	0.53	0.67
Home03	0.16	0.15	0.81	0.82	0.61	0.64
Home04	0.14	0.14	0.62	0.79	0.62	0.79
Home05	0.13	0.12	0.66	0.67	0.56	0.67
Home06	0.10	0.08	0.77	0.82	0.57	0.64
Home07	0.15	0.13	0.67	0.79	0.67	0.64
Home08	0.13	0.11	0.83	0.88	0.63	0.79
Home09	0.16	0.14	0.89	0.91	0.69	0.67
Home10	0.13	0.10	0.69	0.79	0.59	0.64

Table 6. The MAE and RMSE values for the 10 households using the Conv1D and LSTM models.

(1) House ID	Using Conv1D for the Extremely High Power Consumption Records		Using LSTM for the Extremely High Power Consumption Records	
	(2) MAE	(3) RMSE	(4) MAE	(5) RMSE
Home01	1372.46	1517.20	938.18	1113.74
Home02	1209.09	1315.99	872.73	1060.02
Home03	1136.36	1248.64	897.27	1042.41
Home04	1013.64	1147.03	1009.09	1149.90
Home05	1000.82	1198.49	882.64	1073.23
Home06	1081.82	1189.73	781.82	890.35
Home07	1045.46	1192.02	1075.46	1187.65
Home08	1172.73	1313.91	909.09	1059.16
Home09	1354.55	1465.05	1104.55	1261.04
Home10	1081.73	1239.72	940.91	1094.30

5. Conclusions

This research focused on the utilization of smart meter big data, simulating the power consumption forecasting process, and explored other value-added services of smart meters. The main application discussed is predicting whether households are about to experience unsafe electricity events. Therefore, besides energy and cost considerations, the general population is primarily concerned about their own safety. With the completion of smart meter deployments, it is expected that each household's distribution box (i.e., main switch)

will be equipped with smart meters, indicating the possibility of investigating any traces of past electricity usage that may have led to electrical fires. By incorporating the latest prediction technology, early warnings can be provided to households, prompting them to take necessary actions, such as turning off electrical appliances [47,53], thereby reducing the likelihood of fire incidents. This perspective, which focuses on safety, is rarely addressed in the existing literature but formed the core of this study.

Considering that Taiwan has already initiated its smart meter deployment policy, it is believed that improved information security regulations and protection mechanisms related to smart meter big data will contribute to enhancing daily life quality to some extent. This study employed state-of-the-art deep learning algorithms to explore the prediction of unsafe electricity events. It is expected that once databases and servers are installed in each building, the big data analytics of smart meters can be conducted, and this research method can be implemented again. Consequently, TensorFlow-based deep learning applications have gained prominence across various industries in recent years. Deep learning necessitates high-quality big data, and with suitable parameters, it demonstrates predictive capabilities comparable to or even superior to those of humans. The widespread installation of smart meters has become a paramount task for energy conservation and carbon reduction in many countries [54]. However, for smart meters to be truly beneficial, they need to be accompanied by robust information services, such as past electricity consumption forecasting and that proposed in this paper, to incentivize behavior change and reduce the occurrence of residential electrical fires.

In this study, electricity consumption was converted into electricity grade letters, and the parameters of the deep learning model were adjusted to ensure model convergence. The accuracy of the model in the test phase for the 10 households exceeded 93%. In terms of research prospects, the following suggestions are made:

1. Considering the significant generation of big data by smart meters, early planning is crucial for designing data formats and management practices that cater to future information service requirements. While the cost of storage media is decreasing and computing power is improving, it may be more feasible to allow for smaller time intervals between electricity consumption records. This is because once the records have been stored in the format of a longer time interval, it is difficult to transform them into a smaller time interval.

2. Currently, the system developed through this research processes real-world electricity consumption records from smart meters, which entails a large amount of data. However, the database system itself has not been fully optimized, leaving room for improvement in system performance. For instance, this study utilized a relational database without the establishment of an index. Once an index is implemented, it is expected that the system performance will be enhanced. This is because the electricity consumption record data are only added once, while there are multiple query actions in subsequent operations. Considering the infrequent updates to the electricity consumption records, it is advantageous to establish a database index to accelerate the query process.

3. Unsafe electricity usage is a major concern for residents, fire protection agencies, and property insurance providers. The rapid advancements in disaster prevention technologies, smart homes, artificial intelligence, and the Artificial Intelligence of Things (AIoT) are currently hot trends [55]. The early-warning method proposed in this study can be integrated with security and fire departments or residents' personal apps to ensure the safety of lives and property.

4. For residential buildings with unique usage behaviors, such as long-term care institutions or elderly nursing centers, prioritizing the establishment of this early-warning system is advisable. Their electricity consumption patterns are more regular and predictable, leading to more significant disaster avoidance effects [56].

5. Building tax registration data can provide a list of buildings of a certain age. Buildings with aging internal wiring may have a slightly higher fire occurrence rate, making it recommendable to introduce such disaster prevention practices first [57].
6. Issues concerning information security and privacy rights related to smart meter big data should be assigned top priority. The government should expedite the development of local regulations by referencing laws and regulations from advanced countries in Europe and the United States.

References

1. Granderson, J.; Crowe, E.; Touzani, S.; Fernandes, S. *Meter-Based Assessment of the Time and Locational Benefits of a Large Utility's DSM Portfolio*; Research Report 16; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2023.
2. Völker, B.; Reinhardt, A.; Faustine, A.; Pereira, L. Watt's up at Home? Smart Meter Data Analytics from a Consumer-Centric Perspective. *Energies* **2021**, *14*, 719. [CrossRef]
3. Wang, Y.; Chen, Q.; Hong, T.; Kang, C. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Trans. Smart Grid* **2019**, *10*, 3125–3148. [CrossRef]
4. Cai, T.; Dong, M.; Chen, K.; Gong, T. Methods of participating power spot market bidding and settlement for renewable energy systems. *Energy Rep.* **2022**, *8*, 7764–7772. [CrossRef]
5. Schmidt, M.; Åhlund, C. Smart buildings as Cyber-Physical Systems: Data-driven predictive control strategies for energy efficiency. *Renew. Sustain. Energy Rev.* **2018**, *90*, 742–756. [CrossRef]
6. Cano-Ortega, A.; García-Cumbreras, M.A.; Sánchez-Sutil, F.; Hernández, J.C. A Platform for Analysing Huge Amounts of Data from Households, Photovoltaics, and Electrical Vehicles: From Data to Information. *Electronics* **2022**, *11*, 3991. [CrossRef]
7. Stankovic, L.; Stankovic, V.; Liao, J.; Wilson, C. Measuring the energy intensity of domestic activities from smart meter data. *Appl. Energy* **2016**, *183*, 1565–1580. [CrossRef]
8. Liu, X.; Golab, L.; Golab, W.; Ilyas, I.F.; Jin, S. Smart Meter Data Analytics: Systems, Algorithms, and Benchmarking. *ACM Trans. Database Syst.* **2016**, *42*, 2. [CrossRef]
9. Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; Lau, R. Mirror Detection With the Visual Chirality Cue. *IEEE Trans. Patterns Anal. Mach. Intell.* **2023**, *45*, 3492–3504. [CrossRef]
10. Wang, J.Q.; Du, Y.; Wang, J. LSTM based long-term energy consumption prediction with periodicity. *Energy* **2020**, *197*, 117197. [CrossRef]
11. Le, T.; Vo, M.T.; Vo, B.; Hwang, E.; Rho, S.; Baik, S.W. Improving Electric Energy Consumption Prediction Using CNN and Bi-LSTM. *Appl. Sci.* **2019**, *9*, 4237. [CrossRef]
12. Somu, N.; Raman, M.R.G.; Ramamritham, K. A hybrid model for building energy consumption forecasting using long short term memory networks. *Appl. Energy* **2020**, *261*, 114131. [CrossRef]
13. Shen, Y.; Ding, N.; Zheng, H.-T.; Li, Y.; Yang, M. Modeling Relation Paths for Knowledge Graph Completion. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 3607–3617. [CrossRef]
14. Chui, K.T.; Lytras, M.D.; Visvizi, A. Energy Sustainability in Smart Cities: Artificial Intelligence, Smart Monitoring, and Optimization of Energy Consumption. *Energies* **2018**, *11*, 2869. [CrossRef]

15. Athanasiadis, C.; Doukas, D.; Papadopoulos, T.; Chrysopoulos, A. A Scalable Real-Time Non-Intrusive Load Monitoring System for the Estimation of Household Appliance Power Consumption. *Energies* **2021**, *14*, 767. [CrossRef]

16. Chiosa, R.; Piscitelli, M.S.; Capozzoli, A. A Data Analytics-Based Energy Information System (EIS) Tool to Perform Meter-Level Anomaly Detection and Diagnosis in Buildings. *Energies* **2021**, *14*, 237. [CrossRef]

17. Marinakis, V. Big Data for Energy Management and Energy-Efficient Buildings. *Energies* **2020**, *13*, 1555. [CrossRef]

18. Manivannan, M.; Najafi, B.; Rinaldi, F. Machine Learning-Based Short-Term Prediction of Air-Conditioning Load through Smart Meter Analytics. *Energies* **2017**, *10*, 1905. [CrossRef]

19. Ansari, M.H.; Vakili, V.T.; Bahrak, B. Evaluation of big data frameworks for analysis of smart grids. *J. Big Data* **2019**, *6*, 109. [CrossRef]

20. Cerquitelli, T.; Malnati, G.; Apiletti, D. Exploiting Scalable Machine-Learning Distributed Frameworks to Forecast Power Consumption of Buildings. *Energies* **2019**, *12*, 2933. [CrossRef]

21. Wilcox, T.; Jin, N.; Flach, P.; Thumim, J. A Big Data platform for smart meter data analytics. *Comput. Ind.* **2019**, *105*, 250–259. [CrossRef]

22. Toledo-Orozco, M.; Arias-Marin, C.; Álvarez-Bel, C.; Morales-Jadan, D.; Rodríguez-García, J.; Bravo-Padilla, E. Innovative Methodology to Identify Errors in Electric Energy Measurement Systems in Power Utilities. *Energies* **2021**, *14*, 958. [CrossRef]

23. Pritoni, M.; Paine, D.; Fierro, G.; Mosiman, C.; Poplawski, M.; Saha, A.; Bender, J.; Granderson, J. Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis. *Energies* **2021**, *14*, 2024. [CrossRef]

24. Haq, A.U.; Jacobsen, H.-A. Prospects of Appliance-Level Load Monitoring in Off-the-Shelf Energy Monitors: A Technical Review. *Energies* **2018**, *11*, 189. [CrossRef]

25. Sayah, Z.; Kazar, O.; Lejdel, B.; Laouid, A.; Ghenabzia, A. An intelligent system for energy management in smart cities based on big data and ontology. *Smart Sustain. Built Environ.* **2021**, *10*, 169–192. [CrossRef]

26. Apache. Apache Hive. 2023. Available online: https://hive.apache.org (accessed on 1 May 2023).

27. Apache. Apache Spark. 2023. Available online: https://spark.apache.org (accessed on 1 May 2023).

28. Jiang, Z.; Shi, D.; Guo, X.; Xu, G.; Yu, L.; Jing, C. Robust Smart Meter Data Analytics Using Smoothed ALS and Dynamic Time Warping. *Energies* **2018**, *11*, 1401. [CrossRef]

29. Singh, S.; Yassine, A. Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting. *Energies* **2018**, *11*, 452. [CrossRef]

30. Balaji, B.; Bhattacharya, A.; Fierro, G.; Gao, J.; Gluck, J.; Hong, D.; Johansen, A.; Koh, J.; Ploennigs, J.; Agarwal, Y.; et al. Brick: Metadata schema for portable smart building applications. *Appl. Energy* **2018**, *226*, 1273–1292. [CrossRef]

31. Chou, C.C.; Chiang, C.T.; Wu, P.Y.; Chu, C.P.; Lin, C.Y. Spatiotemporal analysis and visualization of power consumption data integrated with building information models for energy savings. *Resour. Conserv. Recycl.* **2017**, *123*, 219–229. [CrossRef]

32. Nguyen, T.A.; Raspitzu, A.; Aiello, M. Ontology-based office activity recognition with applications for energy savings. *J. Ambient Intell. Hum. Comput.* **2013**, *5*, 667–681. [CrossRef]

33. Anvari-Moghaddam, A.; Rahimi-Kian, A.; Mirian, M.S.; Guerrero, J.M. A multi-agent based energy management solution for integrated buildings and microgrid system. *Appl. Energy* **2017**, *203*, 41–56. [CrossRef]

34. Reinhardt, A.; Pereira, L. Special Issue: Energy Data Analytics for Smart Meter Data. *Energies* **2021**, *14*, 5376. [CrossRef]

35. Adams, J.N.; Bélafi, Z.D.; Horváth, M.; Kocsis, J.B.; Csoknyai, T. How Smart Meter Data Analysis Can Support Understanding the Impact of Occupant Behavior on Building Energy Performance: A Comprehensive Review. *Energies* **2021**, *14*, 2502. [CrossRef]

36. Saba, D.; Sahli, Y.; Hadidi, A. An ontology based energy management for smart home. *Sustain. Comput. Inform. Syst.* **2021**, *31*, 100591. [CrossRef]

37. Reda, R.; Carbonaro, A.; de Boer, V.; Siebes, R.; van der Weerdt, R.; Nouwt, B.; Daniele, L. Supporting Smart Home Scenarios Using OWL and SWRL Rules. *Sensors* **2022**, *22*, 4131. [CrossRef]

38. Lork, C.; Choudhary, V.; Hassan, N.U.; Tushar, W.; Yuen, C.; Ng, B.K.K.; Wang, X.; Liu, X. An Ontology-Based Framework for Building Energy Management with IoT. *Electronics* **2019**, *8*, 485. [CrossRef]

39. Bass, B.; New, J.; Ezell, E.; Im, P.; Garrison, E.; Copeland, W. Utility-scale Building Type Assignment Using Smart Meter Data. In Proceedings of the Building Simulation 2021 Conference, Bruges, Belgium, 1–3 September 2021.

40. Ahmadi-Karvigh, S.; Ghahramani, A.; Becerik-Gerber, B.; Soibelman, L. Real-time activity recognition for energy efficiency in buildings. *Appl. Energy* **2018**, *211*, 146–160. [CrossRef]

41. Kofler, M.J.; Reinisch, C.; Kastner, W. A semantic representation of energy-related information in future smart homes. *Energy Build.* **2012**, *47*, 169–179. [CrossRef]

42. Rind, Y.M.; Raza, M.H.; Zubair, M.; Mehmood, M.Q.; Massoud, Y. Smart Energy Meters for Smart Grids, an Internet of Things Perspective. *Energies* **2023**, *16*, 1974. [CrossRef]

43. Hsieh, C.C.; Liu, C.Y.; Wu, P.Y.; Jeng, A.P.; Wang, R.G.; Chou, C.C. Building information modeling services reuse for facility management for semiconductor fabrication plants. *Autom. Constr.* **2019**, *102*, 270–287. [CrossRef]

44. Zhan, S.; Liu, Z.; Chong, A.; Yan, D. Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. *Appl. Energy* **2020**, *269*, 114920. [CrossRef]

45. Bayer, D.; Pruckner, M. A digital twin of a local energy system based on real smart meter data. *Energy Inform.* **2023**, *6*, 8. [CrossRef]

46. Olivares-Rojas, J.C.; Reyes-Archundia, E.; Gutiérrez-Gnecchi, J.A.; González-Murueta, J.W.; Cerda-Jacobo, J. A Multi-Tier Architecture for Data Analytics in Smart Metering Systems. *Simul. Model. Pract. Theory* **2020**, *102*, 102024. [CrossRef]

47. Corno, F.; de Russis, L.; Roffarello, A.M. From Users' Intentions to IF-THEN Rules in the Internet of Things. *ACM Trans. Inf. Syst.* **2021**, *39*, 53. [CrossRef]
48. Park, S.; Ryu, S.; Choi, Y.; Kim, J.; Kim, H. Data-Driven Baseline Estimation of Residential Buildings for Demand Response. *Energies* **2015**, *8*, 10239–10259. [CrossRef]
49. Allen, J.F. Maintaining Knowledge about Temporal Intervals. *Commun. ACM* **1983**, *26*, 832–843. [CrossRef]
50. Sharma, N.; Chakraborty, A.K. Implementation of Dynamic Controls for Grid-Tied-Inverters through Next-Generation Smart Meters and Its Application in Modernized Grid. *Energies* **2022**, *15*, 988. [CrossRef]
51. Lygerakis, F.; Kampelis, N.; Kolokotsa, D. Knowledge Graphs' Ontologies and Applications for Energy Efficiency in Buildings: A Review. *Energies* **2022**, *15*, 7520. [CrossRef]
52. Degha, H.E.; Laallam, F.Z.; Said, B. Intelligent context-awareness system for energy efficiency in smart building based on ontology. *Sustain. Comput. Inform. Syst.* **2019**, *2*, 212–233. [CrossRef]
53. Spoladore, D.; Mahroo, A.; Trombetta, A.; Sacco, M. ComfOnt: A Semantic Framework for Indoor Comfort and Energy Saving In Smart Homes. *Electronics* **2019**, *8*, 1449. [CrossRef]
54. Ang, Y.Q. Using Urban Building Energy Modeling to Develop Carbon Reduction Pathways for Cities. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2022.
55. Chang, C.H.; Chuang, M.L.; Tan, J.C.; Hsieh, C.C.; Chou, C.C. Indoor safety monitoring for falls or restricted areas using Wi-Fi channel state information and deep learning methods in mega building construction projects. *Sustainability* **2022**, *14*, 15034. [CrossRef]
56. Wang, R.G.; Wu, P.Y.; Liu, C.Y.; Tan, J.C.; Chuang, M.L.; Chou, C.C. Route Planning for Fire Rescue Operations in Long-Term Care Facilities Using Ontology and Building Information Models. *Buildings* **2022**, *12*, 1060. [CrossRef]
57. Chou, C.C.; Jeng, A.P.; Chu, C.P.; Chang, C.H.; Wang, R.G. Generation and visualization of earthquake drill scripts for first responders using ontology and serious game platforms. *Adv. Eng. Inform.* **2018**, *38*, 538–554. [CrossRef]