Journal of Electronics, Computer Networking and Applied Mathematics
ISSN: 2799-1156
Vol: 04, No. 03, April-May 2024
http://journal.hmjournals.com/index.php/JECNAM
DOI: https://doi.org/10.55529/jecnam.43.49.60

An Exhaustive and Methodical Analysis of Literature Pertaining To Technigues for Managing Big Data within the Context of the Internet of Things

Dhiaa Mahdi Jaseem[1*], Sefer Kurnaz[2]

[1]Department of Information Technology, IT Faculty, Altinbas University, Istanbul, Turkey.*
[2]Department of Electrical and Computer Engineering, IT Faculty, Altinbas University, Istanbul, Turkey.

Email: [2]sefer.kurnaz@altinbas.edu.tr
Corresponding Email: [1]213721027@ogr.altinbas.edu.tr*

Received: 28 November 2023 **Accepted:** 14 February 2024 **Published:** 01 April 2024

*Abstract: **The merging of big data and the Internet of Things (IoT) has brought both exceptional difficulties and possibilities in data management. This study offers a thorough and methodical examination of the current literature on large data management approaches inside the Internet of Things framework. The study encompasses a broad spectrum of inquiry, ranging from fundamental notions to sophisticated approaches. The Internet of Things (IoT) is a powerful force that seeks to improve user experience and lifestyle. It incorporates several essential technologies including human-machine and machine-to-machine communications, networking technologies, and sensor technologies. Fundamental to the success of the Internet of Things is the effective management of data transmitted through these technologies. This article explores the issues and challenges associated with data management in the context of the Internet of Things and examines different aspects of data, including its sources, collection processes, processing methods, and transmission devices. The article identifies and discusses problems arising from the need to process huge amounts of heterogeneous data in different systems. In relation to these issues, the logical and physical aspects of data management and communications networks are discussed. Additionally, the article takes an in-depth look at the data models used in IoT and explores data management, cleansing, and indexing techniques that take into account the unique characteristics of IoT data. The final sections of the paper comprehensively discuss the benefits and limitations associated with data management in IoT.***

Keywords: Big Data, BD, Big Data and IoT, AI BD, BD Challenges.

1. INTRODUCTION

The advent of the Internet of Things has transformed the way data is generated, collected, and processed, leading to the emergence of Big Data challenges. This paper delves into the interdisciplinary field that explores the intersection of Big Data and IoT, aiming to provide a holistic understanding of the techniques employed to manage the vast volumes of data generated by IoT devices. The realm of technology has been captivated by the phenomenon of the big data, representing the one of the most significant and widespread advancements in the digitals sphere. Its manifestations span across various sources such as the Internet of Things (IoT) devices, smart cities, social networks, and industrial sectors, contributing to an ever-expanding array of data origins. Big data distinguishes itself not solely by its sheer volume but also by its complexity, stemming from the diverse and heterogeneous nature of the information it encompasses[3]. Consequently, big data surpasses the processing capabilities of conventional database systems, whether due to its sheer magnitude, velocity, or its inability to conform to established database structures. Extracting value from such data necessitates alternative processing methods. Merv Adrian's seminal definition characterizes big data as extending beyond the capacities of commonplace hardware and the software tool to capture, manage, and process within acceptable timeframes. Similarly, the McKinsey Global Institute defines it as datasets exceeding the capabilities of conventional database the software tools for captures, storages, managements, and analysis. Essentially, big data emerges as a consequence of data volume outpacing technological capabilities for effective management, storage, and processing[4].

In the realm of electrical energy, a pivotal application area lies in energy prediction and optimization, crucial for maximizing its utilization. Smart cities embody the convergence of various technological strategies aimed at enhancing operational efficiency, service delivery, and overall urban management. By leveraging technologies across transportation, healthcare, education, energy, housing, infrastructure, and environmental sectors, smart cities strive to operate in a cohesive, intelligent manner, fostering increased awareness, interactivity, and effectiveness while also enhancing public access to information and service quality. The Information and Communications Technology (ICT) lies at the core of the development of the smart cities worldwide, aiding in the dissemination and enhancement of sustainable development practices to tackle the challenges posed by escalating urbanization. Governments across many major cities have embraced the smart city concept, initiating the collection of vast datasets to glean valuable insights[5]. Such ideas improve living standards and enhance residents' efforts to achieve sustainable development. The primary factor in increasing the comfort and quality of life of citizens is the need to reduce costs and improve energy consumption. This drive to cut costs has the potential to improve productivity in critical sectors such as education, healthcare, transportation, security, and emergency services. As a result, there is an increasing reliance on big data storage, facilitated by smart grid technologies. For example, monitoring energy consumption in government agencies, such as water and electricity use, has become common practice. The smart grid, a concept intertwined with the Internet of Things, plays a key role in achieving efficient energy consumption, especially in the face of challenges such as population growth and increased energy demand from new equipment[6].

The modernization of electricity grids through smart grid implementation is proposed as a solution. This involves integrating sensors, computers, and communication networks across power generation, transmission, distribution, and load components[7]. Such integration facilitates data acquisition, supply-demand management, and energy usage prediction. Smart meters, for example, enable real-time demand data collection and assist in forecasting future demand levels accurately. These meters, functioning on a two-way communication scheme, provide consumers with detailed insights into their energy consumption while offering additional benefits to utility providers. During peak demand periods, demand response strategies leveraging load behavior and generation knowledge can be deployed. Additionally, network equipment sensors aid in identifying and resolving issues swiftly, thereby enhancing grid reliability and efficiency while minimizing rollout costs. However, the proliferation of such smart grid networks poses new challenges, necessitating the utilization of big data technologies to handle the vast amounts of data generated[8].

The initial embrace of IoT contrasted sharply with today's scene, as illustrated in Figure 1. This gap largely resulted from the steep expenses of computer networks and their limited performance, memory, and storage capacities. In contrast, contemporary IoT devices offer better affordability and feature expanded memory and storage capacities. [9].

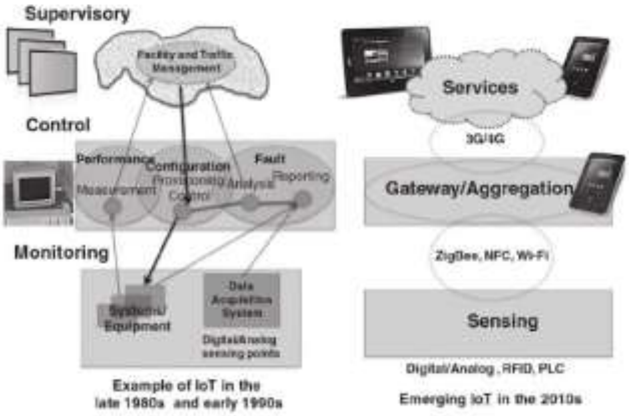

Figure 1 Contrast between Early and Current IoT FunctionalModel

Thanks to these advancements, enterprises can now expand their horizons beyond delivering macro values to also sharing and selling their IoT data on a broader scale. Consequently, we find ourselves in a new era characterized by the rise of service markets, such as big data as a service and insights as a service. Within an IoT environment, three key characteristics stand out: collaboration, event-driven reactivity, and dynamic adaptability (Grabis and Kirikova, 2021). This environment is collaborative, with loosely connected components across distributed devices working together simultaneously to achieve desired functionalities. Moreover, it continuously reacts to various events in physical environments, dynamically establishing connections between devices and collaborations to offer actionable insights[10].

Data in the Internet of Things

Customer information, daily transactions, and organizational operations generate an immense amount of data, often reaching trillions of bytes. This data originates from a multitude of sources, including mobile phones, smart sensors, vehicles, and intelligent equipment. The following overview outlines the methods used to capture, communicate, aggregate, store, and analyze these extensive volumes of data.

In the realm of the Internet of Things (IoT), data serves as the lifeblood driving its functionality and value proposition. Fundamentally, IoT revolves around the concept of connecting various devices to the internet, enabling them to autonomously collect, transmit, and exchange data. This data is sourced from sensors, actuators, and other embedded technologies within these devices. Below is a breakdown of the role and characteristics of data within the IoT ecosystem.:

Data Generation: IoT devices come equipped with sensors designed to collect data from their environments. These sensors have the capability to detect various parameters, including temperature, humidity, motion, light, pressure, and more [11]. The data produced by these sensors offers valuable insights into the physical world, empowering businesses and individuals to monitor, analyze, and react to real-time changes effectively.

Data Transmission: After being gathered, IoT data is transferred over networks, either locally or via the internet, to centralized servers, cloud platforms, or other linked devices. This transmission can take place wirelessly (such as Wi-Fi, Bluetooth, Zigbee) or through wired connections (like Ethernet, Powerline communication). Ensuring efficient data transmission is essential to guarantee timely and dependable communication between IoT devices and the systems responsible for processing and analyzing the data.

Data Processing: The processing of IoT data aims to extract valuable insights and actionable information. This processing takes place at different points along the data journey, including at the edge (i.e., on the device itself), in the cloud, or within hybrid environments. Edge computing specifically entails conducting data processing and analysis directly on the device, which decreases latency, saves bandwidth, and improves privacy and security measures[12].

2. RELATED WORKS

The idea of the 'Internet of Things' (IoT), defined as "sensors and actuators integrated into physical objects connected through wired and wireless networks, often utilizing the same Internet Protocol (IP) that links the Internet," is outlined in [13]. It typically refers to a situation where numerous varied 'things' are linked to the Internet, facilitating communication among them [16]. On the other hand, the definitions of 'Big Data' remain intricate. Although key features involve large volume, diverse variety, and rapid velocity, other characteristics also become crucial, especially when applied to operational processes. The essential features of 'Big Data,' often represented by the classic Vs, are illustrated in Figure 1, with detailed explanations provided in Table 1. Certainly! the " Big Data" refers to the large volumes of structured, semi-structured, and unstructured data that inundate businesses on a day-to-day basis. It's characterized by what's commonly known as the "Three Vs": Volume, Velocity, and Variety. However, over time, additional characteristics have been identified to provide a more comprehensive understanding of Big Data[14].

The term "big data" encompasses the vast, varied, and dynamic scope of data across multiple domains, including its expansion, intricacy, and accessibility. These characteristics often surpass the processing capabilities of standard software applications and traditional database systems. Essentially, big data refers to datasets that exceed the capacities of conventional data processing methods. Traditional analytical systems prove inadequate for handling big data, necessitating the utilization of advanced tools, techniques, and technologies such as classification algorithms, data mining, and platforms like Hadoop and Spark. Key attributes commonly associated with big data include high volume, variety, and velocity, which demand innovative approaches to processing, analyzing, and managing data[15].

In the contexts of the smart sustainable cities, big data typically refers to vast amounts of urban data that pose significant computational and logistical challenges due to their sheer scale and the need for real-time analysis. Urban big data, often labeled with spatial and temporal information, are primarily generated through sensors and automated processes. Although there are no universally accepted definitions of big data, there is a consensus that it holds immense potential for driving advancements and opportunities in various fields in the coming years[16]. Big data is frequently identified by several crucial attributes, commonly known as the Vs, which include volume, variety, and velocity, alongside others like veracity and value. These features underscore the extensive amount, diverse types, and rapid processing speed associated with big data analytics [17].

Big data analytics encompasses the gathering, retention, processing, and interpretation of substantial datasets to extract valuable insights for decision-making. In the realm of smart sustainable cities, big data analytics employs advanced software applications and high-performance computing to derive actionable knowledge across urban sectors such as transportation, environment, and energy. Various types of analytics, including predictive and descriptive, are employed to address specific urban challenges, often relying on techniques such as machine learning and statistical analysis[18].

Urban analytics, rooted in data science principles, focuses on extracting meaningful insights from urban data to address sustainability issues. Techniques such as data mining and regression analysis play a crucial role in understanding urban dynamics and informing decision-making processes. For further exploration of urban analytics and its role in promoting urban sustainability, readers may refer to Bibri (2018).

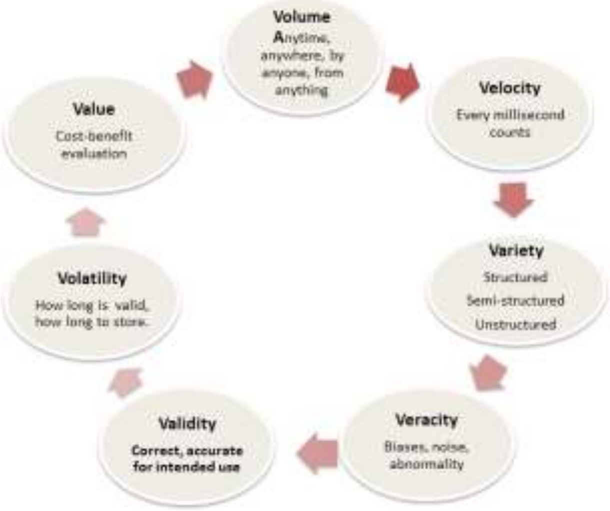
Volume
Anytime, anywhere, by anyone, from anything
Velocity
Every millisecond counts
Variety
Structured
Semi-structured
Unstructured
Veracity
Biases, noise, abnormality
Validity
Correct, accurate for intended use
Volatility
How long is valid, how long to store.
Value
Cost-benefit evaluation

Figure. 2 Definition of Big data as

Table 1. Characteristics of Data in IoT

Characteristics	Description	Example
Volume	Instead of merely capturing business transactions, transporting samples, and aggregating them into another database, modern applications now collect all conceivable data for analysis. These applications read consumption data from numerous sensors connected to various appliances and devices. This reading occurs approximately 100,000 times per day per residence.	anytime, anywhere,by anyone and anything
Velocity	Contemporary applications are collecting streaming data from external systems or sensors, where the data is generated in a continuous manner.	Every millisecond counts
Variety	Data can exist in various formats, including structured, semi-structured, and unstructured types. 1. Structured data possesses semantic meaning and is easily comprehensible for computers.	Database data
	2. Semi-structured data is a variant of structured data that does not adhere to the formal structure typically associated with data models in relational databases.	XML, other mark-up languages
	3. Unstructured data lacks inherent meaning in a manner that a computer cannot interpret its representation.	e-mails, text messages, audio and video streams
Veracity	Addressing biases, noise, and anomalies in data constitutes the most significant challenge when compared to issues of volume and velocity. Essential components in overcoming this challenge include data cleaning processes and measures to safeguard against "dirty data."	dirty data
Validity	If a particular segment of data holds significance or is deemed important, it is imperative to validate the data.	correct and accurate data
Volatility	In certain scenarios, comprehending the available data and its potential duration of storage aids analysts in establishing retention requirements and policies for big data. However, this is not applicable when the data is consistently available for analysis.	Period of data validity and periodof data stored.
Value	This 'value' serves as a critical cost-benefit factor in determining the appropriateness of using 'Big Data,' as its implementation necessitates infrastructure for data gathering, storage, and processing.	cost-benefit evaluation

3. METHODOLOGY

The system functions through a sequence of interconnected and sequential steps, each dependent on the preceding one. Initially, the instantaneous electrical energy consumption is measured and sent from reading devices, like electricity meters, to a designated storage location via the Internet. Subsequently, real-time weather data is obtained and combined with the stored consumption data. Afterward, the data undergoes processing, which includes organization, analysis, and extraction of influential characteristics, to produce results. These results are further examined based on the preprocessing stage. The theoretical framework supporting these system steps is illustrated and explained in Figure 3 below.

A. The Data Preprocessing

The preprocessing methods involves the transformation of raw the data into a comprehensible format. This essential step, known as data preprocessing, constitutes a fundamental aspect of data mining It involves tasks like cleaning and converting data to prepare it for mining operations. Data preparation aims to decrease data volume, recognise data linkages, normalise data, remove outliers, and extract pertinent information.

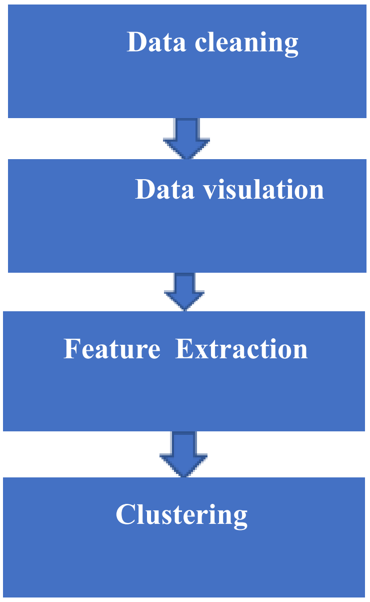

Figure. 3 The Data Preprocessing

This step involves a range of methods including data cleansing, integration, transformation, and reduction [20]. It is an essential preliminary stage that comes before using machine learning or data mining techniques. The main objective is to ensure the quality and format of the data are suitable for the desired analysis.

B. Coefficient of Correlation

Formulas for computing the correlation coefficient are used to evaluate the extent of the link between two characteristics in datasets, as shown in Fig. 4. The computations provide answers within the range of -1 to +1 [21], where:

- A correlation coefficient of +1 signifies a strong positive correlation.
- A correlation coefficient of -1 signifies a strong negative correlation.
- A correlation coefficient of 0 indicates no correlation between the features.

Journal of Electronics, Computer Networking and Applied Mathematics
ISSN: 2799-1156
Vol: 04, No. 03, April-May 2024
http://journal.hmjournals.com/index.php/JECNAM
DOI: https://doi.org/10.55529/jecnam.43.49.60

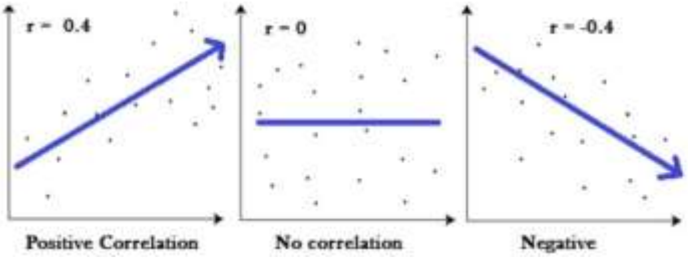

Figure. 4 The Correlation Coefficient Relationships

The magnitude of the correlations coefficients signifies the strength of the association between two feature[22].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

C. Data Visualization

In today's data-driven corporate landscape, visualization plays a pivotal role in transforming abstract data into tangible insights, incorporating elements like length, location, shape, and color. Widely adopted to support decision-making processes, data visualization proves instrumental in providing a comprehensive overview of extensive datasets and aiding data scientists in interpreting analysis results through visual mediums such as the charts, the graphs, and maps [23]. Determining the appropriate type of data visualization forms an integral part of the overall information visualization strategy. There exists a diverse array of visualization options, including scatter charts, line graphs, pie charts, bar charts, heat maps, area charts, choropleth maps, and histograms, as illustrated in Fig. 5

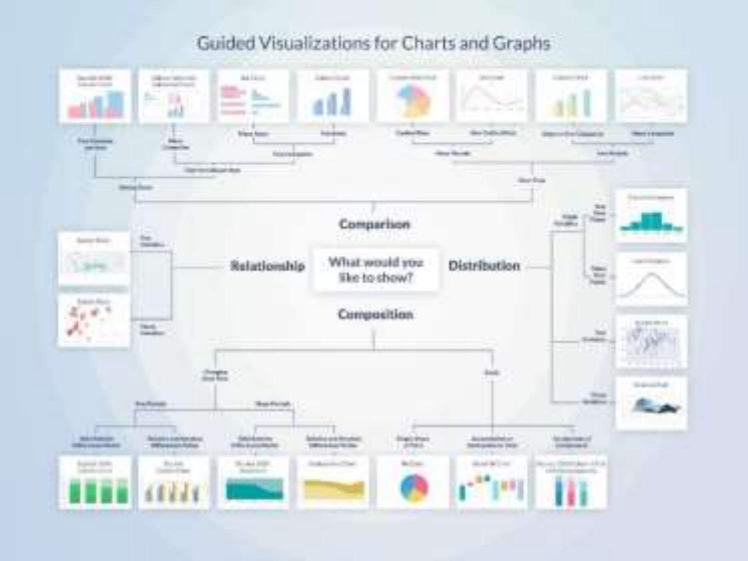
Guided Visualizations for Charts and Graphs
Comparison
Relationship
What would you like to show?
Distribution
Composition

Figure. 5 Visualization Graphs Type

4. RESULTS AND DISCUSSION

A. The Cleaning

Data cleaning encompasses the elimination or substitution of inaccurate data, including the removal of blanks, redundant entries, and noise. Furthermore, it involves populating empty fields with the average daily consumption rate. Then, the daily average electricity consumption for each building is calculated by aggregating the consumption across all buildings and dividing it by the total number of buildings. Figure 4 illustrates the meters collected per day, where the x-coordinate represents the days and the y-coordinate denotes the number of meters recorded on each respective day[24].

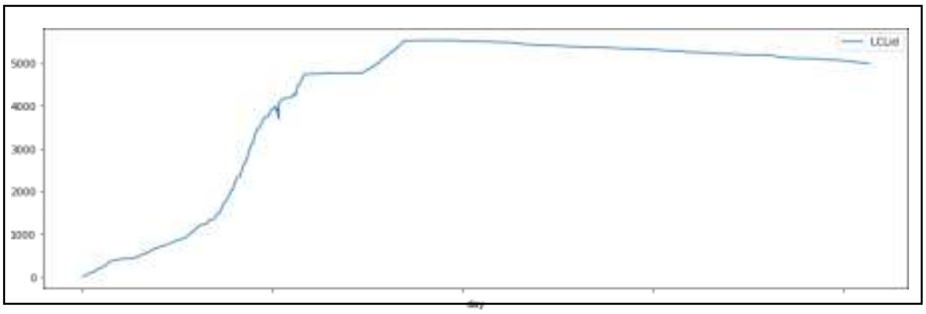

Figure. 6 The Count of the Meters per Day

B. The Visualization

The data visualizations serves as a natural means to provide a comprehensive overview of data and aid in interpreting the analytics results through the visual elements like the charts, graphs, and the maps. It is particularly useful in uncovering relationships between weather conditions and energy consumption for the same time period[25]. Through visualization, the impact of various weather factors on electricity consumption can be observed, elucidating the extent of their influence on consumption percentages:

The Electricity Consumption VS the Temperature

The temperature is one of the key for the weather factors that significantly influences energy consumption, as depicted in Figure 7 below.

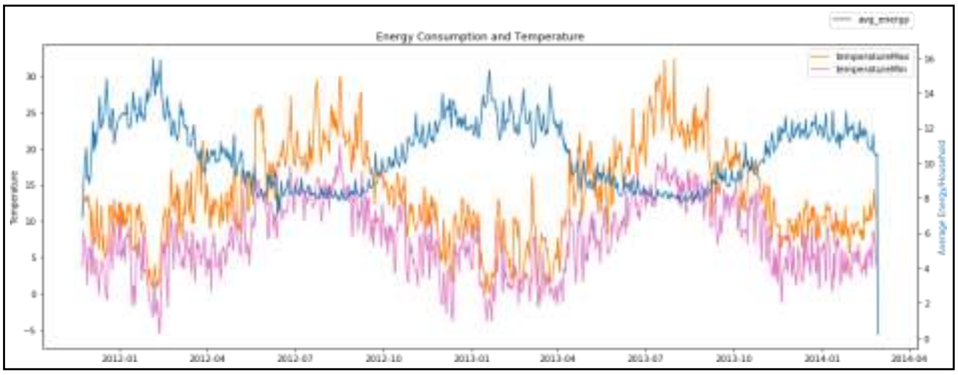

Figure. 7 The Relationship between the Energy and Temperature

The correlation between energy consumption and temperature is inverse, as evidenced by the height of the blue peaks, representing electricity usage, and the dips in the orange (TemperatureMax) and pink (TemperatureMin) regions. This observation aligns with the common understanding that during colder temperatures, energy consumption typically increases, likely attributable to higher usage of heaters to counter the cold weather[26].

Humidity

Humidity is another weather factors that does not exert the significant impact on energy consumption. The relationship between humidity and average energy consumption shows no clear direction, indicating a lack of distinct correlation between them. Regardless of whether the humidity is high or low, energy consumption in a humid environment remains consistent, as depicted by the colors indicating the same direction in Figure 8 below:

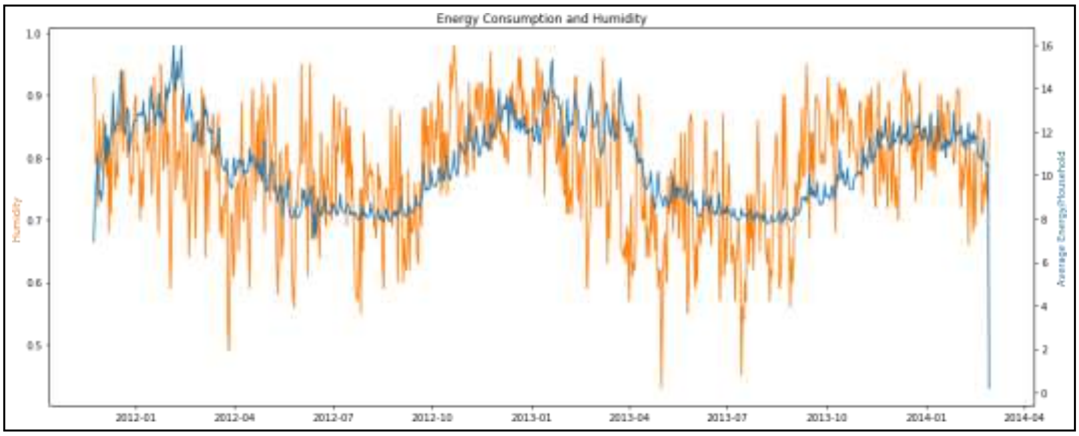

Figure. 8 The Relationship between the Energy and Humidity

5. CONCLUSION

This study outlines a methodology for collecting, analyzing, and visualizing the relationship between selected weather factors (such as temperature, dew point, and UV index) and energy consumption values derived from electrical meters utilizing the Internet of Things. The research focuses on gathering energy consumption data for buildings within a specific area and computing the total consumption rate for the entire area. Subsequently, the data undergo preprocessing, which involves several stages including cleaning (eliminating duplicate, damaged, inaccurate, or incomplete data), visualization (converting abstract data into concrete insights), feature extraction (identifying the most significant features), and clustering (segmenting the data into clusters based on these features using the k-means model).

6. REFERENCES

1. EIOPA (2021). Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in the European Insurance Sector (European Insurance and Occupational Pensions Authority).
2. EIOPA (2019). Big Data Analytics in Motor and Health Insurance: A Thematic Review

(European Insurance and Occupational Pensions Authority).

3. Agrawal, A., Gans, J.S., and Goldfarb, A. (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. J. Econ. Perspect. 33, 31–50.
4. Jang, J.A., Kim, H.S., and Cho, H.B. (2011). Smart roadside system for driver assistance and safety warnings: framework and applications. Sensors 11, 7420–7436.
5. Masters, J. (2012). Telematics Making Fresh Gains, 18 (ITS International).
6. Shannon, D., Jannusch, T., David-Spickermann, F., Mullins, M., Cunneen, M., and Murphy, F. (2021). Connected and autonomous vehicle injury loss events: potential risk and actuarial considerations for primary insurers. Risk Manag. Insur. Rev. 24, 5–35.
7. Titsworth, T. (2002). Telematics might steer your car into the future. IEEE Multimedia 9, 9–11.
8. Cie´slik, B. (2017). Telematics in automobile insurance. Collegium Econ. Anal. Ann. 45, 79–92.
9. Constantinescu, C.C., Stancu, I., and Panait, I. (2018). Impact study of telematics auto insurance. Rev. Financ. Stud. 3, 17–35.
10. Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., et al. (2014). Insurance telematics: opportunities and challenges with the smartphone solution. IEEE Intell. Transp. Syst. Mag. 6, 57–70.
11. Husnjak, S., Perakovic´, D., Forenbacher, I., and Mumdziev, M. (2015). Telematics system in usage based motor insurance. Proced. Eng. 100, 816–825.
12. Yoon, D., Choi, J., Kim, H., and Kim, J. (2008). Future automotive insur- ance system based on telematics technology. In 2008 10th International Conference on Advanced Communication Technology (IEEE).
13. Cunneen, M., and Mullins, M. (2019). Framing risk, the new phenomenon of data surveillance and data monetisation; from an 'always-on' culture to 'always-on' artificial intelligence assistants (Hybrid Worlds), p. 65.
14. Cunneen, M., Mullins, M., and Murphy, F. (2019). Artificial intelligence as- sistants and risk: framing a connectivity risk narrative. AI Soc. 35, 625–634.
15. Cunneen, M., Mullins, M., Murphy, F., Shannon, D., Furxhi, I., and Ryan,
16. C. (2020). Autonomous vehicles and avoiding the trolley (dilemma): vehicle perception, classification, and the challenges of framing decision ethics. Cybernetics Syst. 51, 59–80.
17. van den Boom, F. (2021). Regulating Telematics Insurance. Insurance Distribution Directive (Cham: Springer), pp. 293–325.
18. Irode, P.L. (2017). An Auto Telematics System for Insurance Premium Rating & Pricing (University of Nairobi).
19. Sreethamol, P., Sulfiya, K., and Vineeth, K. (2018). Consumers percep- tion towards telematics in insurance. Res. J. Human. Soc. Sci. 9, 657–662.
20. Kuo, K., and Lupton, D. (2020). Towards explainability of machine learning models in insurance pricing. arXiv arXiv:200310674.
21. Ma, Y.-L., Zhu, X., Hu, X., and Chiu, Y.-C. (2018). The use of context-sen- sitive insurance telematics data in auto insurance rate making. Transportation Res. A: Policy Pract. 113, 243–258.
22. Hollis, A., and Strauss, J. (2007). Privacy, Driving Data and Automobile Insurance: An Economic Analysis (University Library of Munich: Germany.: MPRA Paper 11091).

23. Filipova-Neumann, L., and Welzel, P. (2010). Reducing asymmetric infor- mation in insurance markets: cars with black boxes. Telematics Inform. 27, 394–403.
24. Kargupta, H. (2012). Connected cars: how distributed data mining is changing the next generation of vehicle telematics products. In International Conference on Sensor Systems and Software, F. Martins and H. Paulino, eds. (Berlin, Heidelberg: Springer).
25. Cunneen, M., Mullins, M., Murphy, F., and Gaines, S. (2019). Artificial driving intelligence and moral agency: examining the decision ontology of unavoidable road traffic accidents through the prism of the trolley dilemma. Appl. Artif. Intell. 33, 267–293.
26. Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting (Oxford University Press).
27. Mizgier, K.J., Kocsis, O., and Wagner, S.M. (2018). Zurich Insurance uses data analytics to leverage the BI insurance proposition. Interfaces 48, 94–107.