



DELIVERABLE 1.3

Scientific report on IoT and digital twins in smart buildings

31 03 2025



Funded by
the European Union

DOCUMENT TRACKING DETAILS

Project acronym	SmartWins
Project title	Boosting Research for a Smart and Carbon Neutral Built Environment with Digital Twins
Starting date	01/10/2022
Duration	36 months
Call identifier	Twinning (HORIZON-WIDERA-2021-ACCESS-03)
Grant agreement number	101078997

Deliverable information	
Deliverable number	1.3
Work package number	1
Deliverable title	Scientific report on IoT and Digital Twins in smart buildings
Lead beneficiary	KTU
Authors	Nikos Tsalikidis (CERTH) Lina Morkūnaitė (KTU) Paulius Spūdys (KTU) Paraskevas Koukaras (CERTH) Paris Fokaides (KTU)
Due date	31/03/25
Actual submission date	04/04/25
Type of deliverable	R
Dissemination level	SEN

Legal Disclaimer

The SmartWins project funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union

or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

VERSION MANAGEMENT

Revision table

Version	Name	Date	Description
0.1	Nikos Tsalikidis	20/01/2025	TOC
0.2	Nikos Tsalikidis, Paraskevas Koukaras	05/02/2025	Revised TOC, conceptualization, mapping of subchapters
0.5	Nikos Tsalikidis, Paraskevas Koukaras	28/02/2025	Initial draft full draft
0.7	Nikos Tsalikidis, Paraskevas Koukaras KTU team	16/03/2025	CERTH components full developed. Communicated draft to KTU
0.9	Nikos Tsalikidis, Paraskevas Koukaras KTU team	31/03/2025	Full developed, finalised draft

EXECUTIVE SUMMARY

Deliverable 1.3 of the SmartWins project presents a comprehensive scientific report on the integration of Internet of Things (IoT) technologies and Digital Twins (DTs) in smart buildings. The report highlights the collaboration between CERTH and KTU, focusing on the knowledge transfer, technical training, and joint research activities that enabled the development of advanced tools for energy forecasting, system optimization, and real-time performance assessment in buildings and urban environments.

The report outlines a structured training program delivered by CERTH, consisting of two rounds of seminars—first on foundational topics such as data storage, transmission, and interoperability, and second on advanced applications including machine learning (ML), neural networks, and ensemble forecasting. These sessions significantly enhanced the technical capacities of KTU researchers in processing, analyzing, and integrating IoT data with DT frameworks.

Two joint scientific publications emerged from this collaboration. The first developed a 24-hour predictive model for urban traffic congestion using multi-source sensor data and weather inputs, while the second analyzed hot-water energy consumption in residential buildings under crisis scenarios using a hybrid ensemble classification model.

Experimental designs incorporated real-world datasets and advanced ML algorithms to validate data-driven control strategies for energy and mobility systems. The findings demonstrate that integrating DTs with IoT and AI can greatly enhance operational efficiency, crisis responsiveness, and sustainability in the built environment.

TABLE OF CONTENT

1. Introduction	10
1.1. Scope and objectives	10
1.2. Relation to other Tasks and Deliverables	10
2. Background	10
2.1. Digital Twins, Smart Buildings, Smart Cities	11
2.2. IoT-Driven Data in the Built Environment and Beyond	11
2.3. Forecasting Energy Performance with Machine Learning	12
3. Gained Knowledge by KTU	13
3.1. Online seminars & training sessions from CERTH	13
3.1.1. Objectives and structure	13
3.1.1. Round 1: Foundational Topics	14
3.1.1.1. Seminar 1- Data analysis	14
3.1.1.2. Seminar 2 - Data storage	15
3.1.1.3. Seminar 3 - Data integration & interoperability	16
3.1.1.4. Seminar 4 - Data transmission	18
3.1.1.5. Seminar 5 - Infrastructure for Big Data management	20
3.1.1. Round 2: Advanced Applications and Techniques	21
3.1.1.1. Seminar 1: Advanced Data Processing and Feature Engineering	21
3.1.1.2. Seminar 2: Advanced Machine Learning for Time Series Energy Forecasting	22
3.1.1.3. Seminar 3: Artificial Neural Networks and Multi-Model Forecasting	23
3.1.2. Follow-up discussions and progress tracking	24
4. Joint research activities and achievements	25
4.1. Preparatory phase	25
4.1.1. Planning, Tools and Resources for collaboration	25
4.1.2. Identification of relevant research topics	27
4.2. Experimental design and implementation	29
4.2.1. Publication 1	31

4.2.2. Publication 2	35
4.3. Results Analysis and Insights	40
4.3.1. Publication 1	40
4.3.2. Publication 2	42
4.1. Implications and final publication	45
5. Overview of KTU gains from collaboration with CERTH	47
5.1. Technical skills	47
5.2. Machine Learning methods application	52
5.3. Multivariate data analysis methods application	53
6. Conclusions	55
References	59

List of Figures

Figure 1. Overview of asset core engine of 'D ² EPC' project.....	14
--	----



Figure 2. Example of a building energy load forecasting strategy.....	15
Figure 3. Example of a JSON file energy load forecasting strategy.....	16
Figure 4. A complete system architecture of a Digital Twin infrastructure.....	17
Figure 5. Indicative example of an IFC schema.....	17
Figure 6. Hardware installed in the CERTH nZEB Smart House.....	19
Figure 7. nZEB Smart House IoT Platform	19
Figure 8. DT & Big data fusion example	20
Figure 9. Locations of measured traffic flow data in the city of Trondheim.	31
Figure 10 Weather data feature selection	32
Figure 11. Overview of forecasting approach.	34
Figure 12. Arrangement of case study buildings in Kaunas, Lithuania.	35
Figure 13. Mobility changes in relation to crisis severity levels.....	37
Figure 14. Overall methodological approach.....	39
Figure 15. 48h vehicle count pattern for selected traffic locations.	41
Figure 16. Direct multi-step forecasting (R-squared comparison).....	42
Figure 17. Daily profiles during different periods.	43
Figure 18. Extracted daily profiles based on crisis severity level.....	44

List of Tables

Table 1. Overview of planned activity.....	26
Table 2. Overview of focus areas derived from literature	28

1. Introduction

1.1. Scope and objectives

The growing demand for energy-efficient, sustainable buildings has accelerated the adoption of smart technologies such as IoT and DTs. These technologies provide the tools needed to monitor, analyze, and optimize energy consumption in real time, offering both environmental and economic benefits.

The primary objective of this deliverable is to establish a framework that integrates IoT-enabled data acquisition with predictive analytics powered by DTs. This deliverable contributes to Task 1.3 by focusing on monitoring, modeling, and optimizing energy performance of buildings.

The scope of this deliverable encompasses:

- The application of IoT technologies to collect and process real-time building performance data.
- The development of methodologies for energy performance forecasting using advanced machine learning (ML) techniques.
- The integration of DTs with IoT to create dynamic, data-driven models that facilitate actionable insights.
- Compliance with standards such as openBIM to ensure data interoperability and seamless integration across platforms.

1.2. Relation to other Tasks and Deliverables

- Identify dependencies on other tasks.
- Describe how this deliverable contributes to broader project milestones.

2. Background

The transition towards a carbon-neutral built environment necessitates innovative approaches for the design, monitoring, and optimization of building energy systems. The integration of Internet of Things (IoT) technologies and Digital Twins (DTs) offers a transformative solution for achieving this vision. By creating dynamic, data-driven models of physical assets, DTs enable real-time monitoring and predictive analytics, providing opportunities for optimizing energy consumption and enhancing operational efficiency.

2.1. Digital Twins, Smart Buildings, Smart Cities

Digital Twins(DT) are a cornerstone of smart building technologies, linking physical components (e.g., HVAC systems, energy meters) with their digital counterparts. Through bidirectional data exchange, DTs enable real-time simulation, fault detection, scenario analysis, and optimization, allowing stakeholders to evaluate system performance under varying conditions.

- DTs leverage Building Information Modeling (BIM) standards, such as Industry Foundation Classes (IFC) and Green Building XML (gbXML), to ensure semantic interoperability across diverse platforms and domains.
- By incorporating live data feeds from IoT networks, digital twins can monitor indoor climate, energy use, and occupant behavior, supporting applications such as demand-response, predictive maintenance, and thermal comfort optimization.
- At the urban scale, city-level digital twins extend these capabilities to infrastructure systems (e.g., energy, water, transportation), facilitating integrated urban planning and resource optimization aligned with sustainability and resilience goals.

2.2. IoT-Driven Data in the Built Environment and Beyond

IoT technologies serve as the foundation for real-time data acquisition in smart buildings and cities. Distributed networks of sensors and actuators provide granular, time-resolved data on critical operational metrics:

- Energy consumption, thermal loads, lighting usage
- Indoor environmental quality, including temperature, humidity, and CO₂ levels
- Occupancy and behavior patterns, supporting user-centric adaptive control

These data streams are transmitted via standard communication protocols (e.g., ZigBee, Modbus, BACnet) and integrated into Building Management Systems (BMS) or cloud-based platforms that support edge analytics and system coordination.

In the broader smart city context, IoT enables applications in Intelligent Transportation Systems (ITS), urban mobility monitoring, and environmental sensing. For instance, real-time traffic flow data, vehicle counts, and weather conditions are collected to optimize signal timing, reduce congestion, and enhance air quality. These systems

benefit from the same digital twin principles applied in buildings, offering a unified framework for urban-scale decision support.

Moreover, the interoperability enabled by openBIM, semantic web technologies, and linked data approaches enhances the integration of building and city systems, setting the stage for cross-domain coordination—from energy optimization to climate adaptation.

2.3. Forecasting Energy Performance with Machine Learning

The use of machine learning (ML) in forecasting the energy performance of buildings represents a significant advancement in achieving sustainability goals. ML models can analyze historical and real-time data to predict energy consumption, generation, and efficiency with high accuracy. By incorporating factors such as weather patterns, occupancy rates, and building system usage, ML-driven forecasting enables proactive energy management and optimization.

Machine learning (ML) provides powerful tools for forecasting the energy performance of buildings, enabling more efficient, adaptive, and data-informed management strategies. By analyzing historical trends and real-time sensor data, ML models can uncover patterns and generate predictions related to energy consumption, system performance, and operational behavior. Key applications of ML in this context include:

- Short- and long-term forecasting of energy demand (e.g., electricity, heating, cooling).
- Detection of consumption anomalies or abnormal usage patterns.
- Support for predictive control strategies, such as adjusting HVAC settings based on expected occupancy or weather changes.

A variety of algorithms can be applied, from traditional regression models to advanced methods such as decision trees, ensemble learners (e.g., gradient boosting), and sequence-based models (e.g., recurrent neural networks).

These predictive models are often integrated into digital twin environments, where forecasts can inform simulations of future building states and support decision-making processes. This enables a shift from reactive to proactive building operation, improving overall performance and resilience.

The benefits of ML-based forecasting include:

- Enhanced operational efficiency, through better alignment of energy supply and demand.
- Reduced energy waste, via informed scheduling and load management.
- Support for renewable integration, enabling more intelligent use of on-site generation and storage.

By enabling data-driven forecasting and adaptive control, ML plays a critical role in the evolution of intelligent building systems and contributes directly to the goals of energy efficiency, comfort, and sustainability.

3. Gained Knowledge by KTU

3.1. Online seminars & training sessions from CERTH

3.1.1. Objectives and structure

The knowledge transfer seminars conducted by CERTH were designed to enhance the technical expertise of project partners in deploying IoT and DT technologies. These seminars served as a platform for introducing advanced methodologies and tools for data acquisition, big data management, and predictive analytics. The training was divided into two rounds:

A. Round 1: Introductory seminars on key topics regarding IoT and Digital Twin applications

The first round of training focused on fundamental concepts necessary for deploying IoT and DT technologies. These sessions covered a broad spectrum of essential topics, ensuring all participants were equipped with baseline knowledge for further specialization:

- a) Data analysis: structure, formats, taxonomies, available ontologies.
- b) Data storage
- c) Data integration & interoperability
- d) Data transmission
- e) Big Data management

B. Round 2: Advanced Applications and Techniques

The second round of training sessions focused on equipping KTU members with advanced technical expertise. These sessions emphasized hands-on applications and specialized methodologies for addressing real-world

challenges. All seminars were conducted online, and accompanying code and practical examples were presented and shared with KTU to facilitate further experimentation and application.

Seminar Topics:

- a) Seminar 1: Data Processing and Feature Engineering
- b) Seminar 2: Advanced Machine Learning for Time Series Energy Forecasting
- c) Seminar 3: Artificial Neural Networks and Multi-Model Forecasting

3.1.1. Round 1: Foundational Topics

3.1.1.1. Seminar 1- Data analysis

This seminar focused on techniques for preprocessing and analyzing building performance data. Participants learned about taxonomy-based structuring methods and available ontologies, such as those aligning with ISO standards for energy performance assessment. Emphasis was placed on real-time visualization of building information to support decision-making and predictive modeling.

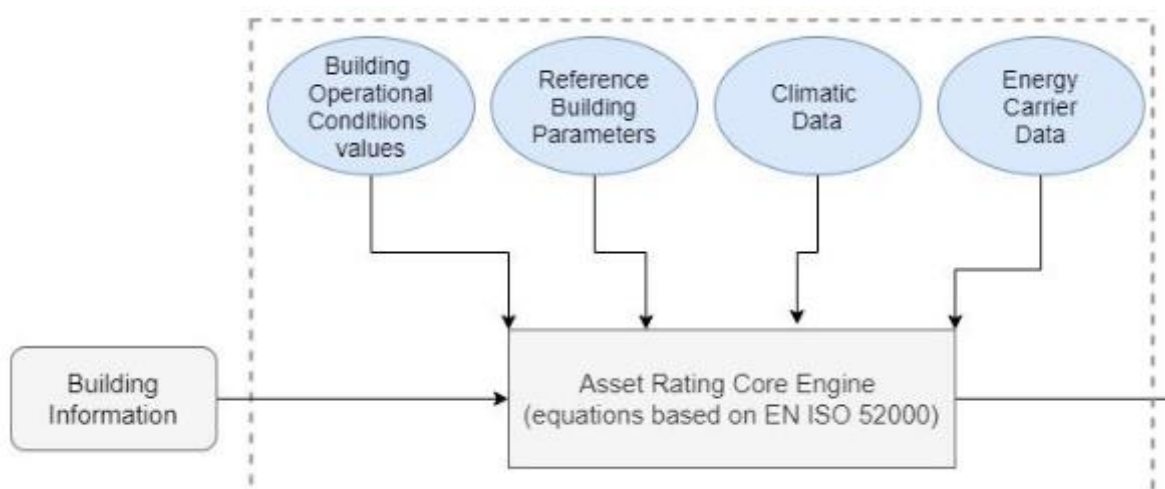


Figure 1. Overview of asset core engine of 'D^2EPC' project

ML plays a pivotal role in improving the accuracy of energy forecasting within the broader scope of smart building management. By leveraging historical data and real-time IoT inputs, ML models can predict key parameters such as electrical energy

demand, heating and cooling loads, and renewable energy generation. Effective data pre-processing ensures the reliability and accuracy of forecasting models

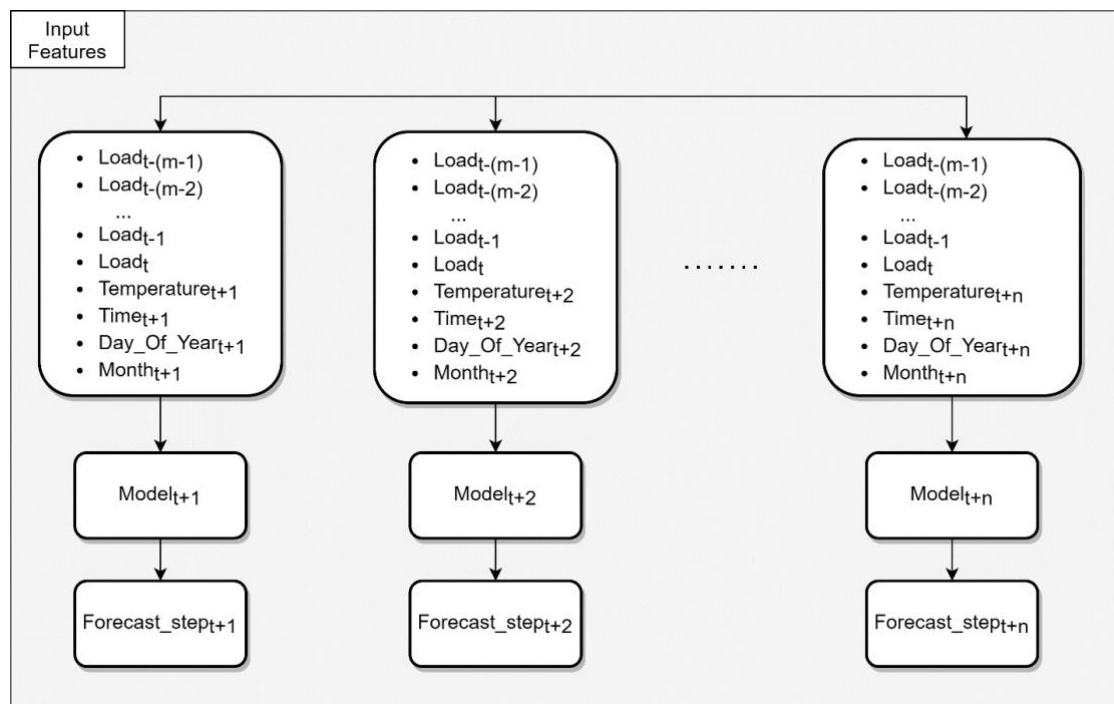


Figure 2. Example of a building energy load forecasting strategy

When integrated with IoT platforms and DT technologies, these methodologies enable real-time optimization of building systems, leading to enhanced operational efficiency. Key applications include energy load forecasting and management, as well as the seamless integration of renewable energy production systems, ensuring a sustainable and responsive approach to smart building performance.

3.1.1.2. Seminar 2 - Data storage

Participants explored database architectures suitable for IoT data, including relational databases like PostgreSQL and non-relational databases such as MongoDB. The seminar covered data modeling techniques, schema design, and the role of JSON for maintaining relationships between data entities.

In IoT-enabled smart buildings typically a two-phase data storing and data relationship process:

1. **Physical Data Model:** Organized within a DBMS schema, where tables represent entities like buildings, systems, or sensors.
2. **JSON Files:** Used for data validation and storage, describing attributes and types without requiring specific query languages. Relationships are managed via primary and foreign keys.

```
{
  "building": {
    "building_id": {
      "datatype": "string",
      "value": "A string value",
      "description": "A unique string defining a building ID - Primary Key for Buildings Table"
    },
    "location_id": {
      "datatype": "string",
      "value": "A string value",
      "description": "A unique string defining a location ID - Foreign Key to Locations Table"
    },
    "use_id": {
      "datatype": "string",
      "value": "A string value",
      "description": "A string describing briefly the usage of a building"
    },
    "ownership": {
      "datatype": "string",
      "value": "A string value",
      "description": "Valuable information about building ownership"
    }
  }
}
```

Figure 3. Example of a JSON file energy load forecasting strategy

3.1.1.3. Seminar 3 - Data integration & interoperability

The focus of this seminar was on integrating IoT data with digital twins using open standards like openBIM to ensure interoperability between diverse systems.

1 openBIM Approach:

- Encourages collaboration using open standards to address limitations of proprietary technologies.
- Enables bi-directional data flow between Building Information Modeling (BIM) databases and digital twin systems for enhanced integration.

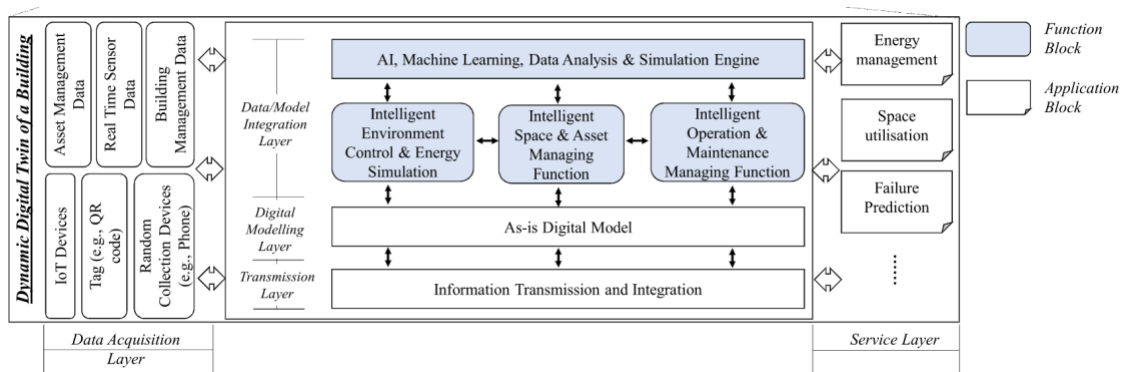


Figure 4. A complete system architecture of a Digital Twin infrastructure

2 Key Standards:

- **Industry Foundation Classes (IFC):**

- Provides an object-oriented, hierarchical schema for data interoperability, classification, and visualization.
- Supports complex geometry representation and semantic data enrichment.
- APIs such as IfcOpenShell and BIMserver enable querying and modification of IFC data.

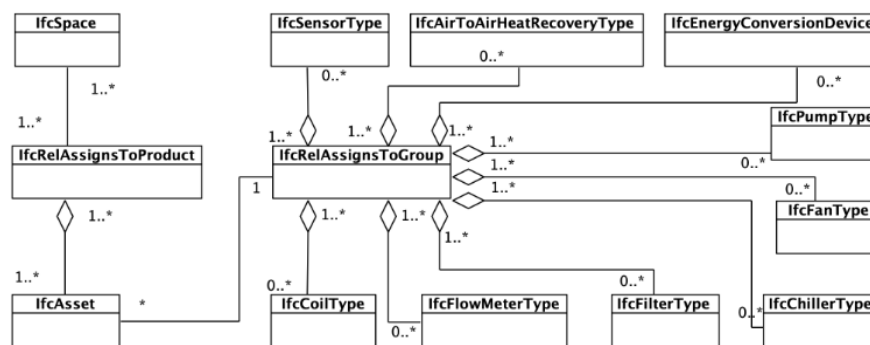


Figure 5. Indicative example of an IFC schema

- **Green Building XML (gbXML):**

- Facilitates the transfer of building information between architectural and engineering models.

- Developed in XML format, it ensures compatibility across platforms with minimal human intervention.

The framework facilitates real-time data exchange and system interaction, enabling advanced decision-making and optimization processes.

This seminar delved into the integration of IoT data streams with Digital Twins using open standards like IFC and gbXML. By employing APIs like IfcOpenShell and BIMserver, participants explored how to establish bi-directional communication between physical sensors and digital representations.

3.1.1.4. Seminar 4 - Data transmission

Participants were introduced to IoT communication protocols. These include wired protocols such as BACnet, Modbus, and Canbus, as well as wireless protocols like ZigBee, EnOcean, Z-Wave, BLE, and LoRa. including MQTT, ZigBee, and LoRaWAN, for transmitting real-time data.

The session also emphasized the role of gateways in managing data flows between sensors and cloud-based Building Management Systems (BMS).

CERTH Smart House IoT Infrastructure

The CERTH nZEB Smart House infrastructure exemplifies the integration of IoT for smart building applications, facilitating real-time monitoring and diverse testing opportunities. This infrastructure communicates through multiple protocols and gateways, ultimately converging data into an IoT platform. The platform serves as the core for data collection, handling, and interaction through well-defined interfaces. Its IoT platform collects and processes data via:

- **Wired Protocols:** Includes industry standards such as BACnet, Modbus, and Canbus.
- **Wireless Protocols:** Supports technologies like ZigBee, EnOcean, Z-Wave, BLE, and LoRa.
- **Ecosystem Diversity:** These protocols create a robust and versatile ecosystem for testing scenarios, enhancing the infrastructure's capacity to adapt to various requirements.





Type of device	ID number	Image	Location	Connectivity	Data Access	Measuring interval
TEMPERATURE & HUMIDITY						
Plugwise Sense	-		One per each space in the main House	ZigBee HA 2.0	SmartHome IoT Platform API	15' -adjustable
Gavazzi EM340	-		One per indoor electric panel (3 ground floor)	Modbus RTU	SmartHome IoT Platform API	1 sec from device 1 min from API
Gavazzi EM270	-		Building Point of Common Coupling	Modbus RTU	SmartHome IoT Platform API	1 sec from device 1 min from API
Thermokon SR04 CO2	-		1 living room ground floor and 1 playroom first floor	EnOcean		Measuring interval WakeUp time = 100 sec. (default) Transmission interval every 100 sec. at change >0,4 K, >2,5% rH or 50 ppm, otherwise every 1000 sec.

Figure 6. Hardware installed in the CERTH nZEB Smart House



Figure 7. nZEB Smart House IoT Platform

3.1.1.5. Seminar 5 - Infrastructure for Big Data management

This seminar provided insights into big data infrastructures for processing IoT sensor data. Participants learned to utilize Hadoop Distributed File Systems (HDFS) for scalable data storage and Apache Spark for real-time data analytics. A reference architecture was described integrating advanced big data technologies to ensure effective data ingestion, processing, and visualization, facilitating seamless management of building systems and services. Steps in the Big Data Architecture:

- **Data Storage:** IoT sensor data is stored in the **HDFS**
- **Data Ingestion:** **Apache Flume** ingests sensor data into HDFS, ensuring fault tolerance, failover, and recovery. Flume agents use IoT data as the source and HDFS as the destination (sink).
- **Real-Time Analytics:** Data in HDFS is processed using **Apache Spark**, providing near-real-time insights for decision-making and managing building devices effectively.
- **Visualization:** Tools such as **Kibana** and **Power BI** present near-real-time summaries for actionable reporting and monitoring.

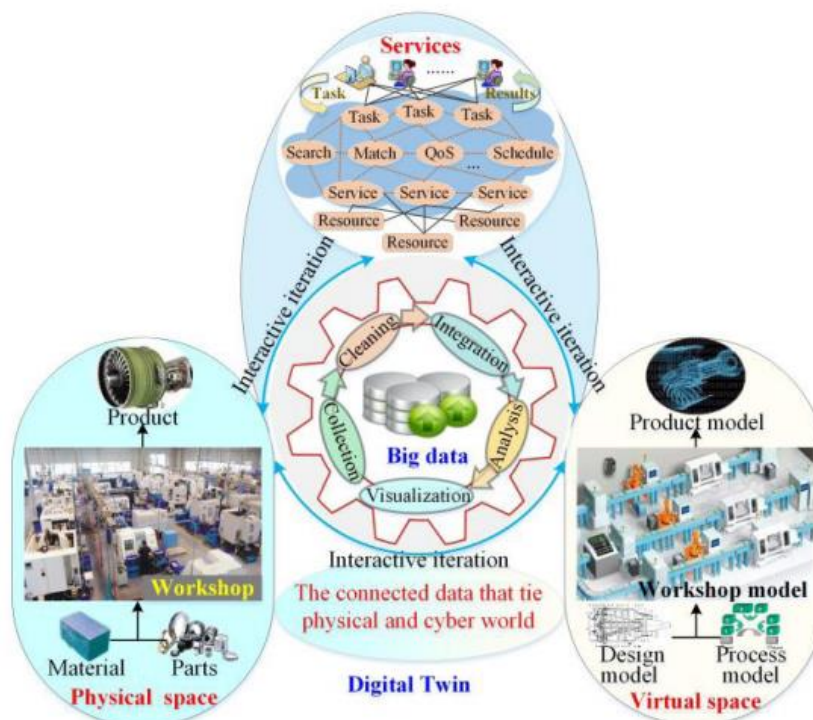


Figure 8. DT & Big data fusion example

3.1.1. Round 2: Advanced Applications and Techniques

3.1.1.1. Seminar 1: Advanced Data Processing and Feature Engineering

This seminar focused on highlighting the importance of feature selection, data scaling, and advanced data processing to optimize input data for machine learning models.

Key Topics Covered:

1. Feature Selection and Correlation Analysis:

- Participants learned to identify highly correlated features using correlation matrices, which help reduce redundancy in datasets and improve model interpretability.
- The significance of selecting meaningful features for accurate predictions was demonstrated using visual tools and statistical measures.

2. Data Scaling and Transformation:

- Techniques such as Min-Max scaling, standardization, and logarithmic transformations were discussed to ensure consistent input ranges across variables.
- These methods enhance model performance, particularly for algorithms sensitive to feature magnitudes.

3. Synthetic Feature Generation:

- Participants were introduced to creating synthetic features, including statistical aggregations (e.g., mean, median, rolling averages) and time-based features (e.g., day-of-week, holiday markers).
- The seminar highlighted how these features enrich the predictive power of datasets, enabling models to capture temporal and seasonal patterns effectively.

4. Data Cleaning and Outlier Management:

- Practical examples showed how to identify and handle missing data and outliers using interpolation and percentile-based thresholds.
- Participants explored strategies to maintain data integrity while ensuring minimal loss of information.

After the above topics were presented, through hands-on exercises, participants:

- Applied **data scaling** and **feature selection** techniques to building energy datasets.
- Enhanced datasets by engineering synthetic features, integrating external factors like holidays and working hours.
- Understood the role of data cleaning in addressing outliers and ensuring consistency.
- Evaluated machine learning models using metrics such as **MAPE**, **RMSE**, and **R²**, gaining practical insights into predictive performance.
- Experimenting with correlation heatmaps and scatter plots to identify relationships between variables.

3.1.1.2. Seminar 2: Advanced Machine Learning for Time Series Energy Forecasting

This session focused on advanced ensemble learning methods and machine learning techniques for energy forecasting in smart buildings. Participants gained hands-on experience with state-of-the-art algorithms and explored optimization strategies.

Key Topics Covered

1 Application of Machine Learning Algorithm

A variety of machine learning algorithms were applied to energy forecasting problems, including **LightGBM**, **XGBoost**, and **CatBoost**.

2 Hyperparameter Tuning:

Participants were introduced to optimization techniques for improving model performance:

- **GridSearchCV:** Systematically evaluates combinations of hyperparameters across a predefined grid.
- **RandomizedSearchCV:** Samples random combinations of hyperparameters for quicker tuning when computational resources are limited.

3 Introduction to Ensemble Learning:

Participants explored ensemble learning, a powerful technique where multiple models are combined to enhance predictive performance

- **Voting Regressor:** Combined predictions from multiple algorithms (e.g., Random Forest, Support Vector Regression, and LassoCV) using both hard and soft voting methods to improve overall accuracy.
- **Stacking Ensemble:** Introduced a meta-model that integrates base model predictions to optimize forecasting results. This hierarchical structure was particularly effective in capturing complex patterns.

3.1.1.3. Seminar 3: Artificial Neural Networks and Multi-Model Forecasting

This seminar delved into advanced forecasting methodologies with an emphasis on artificial neural networks (ANNs) and multi-step time series predictions. Participants were introduced to the design and application of deep learning models and hybrid approaches to enhance forecasting capabilities.

Key Topics Covered

1 Multi-Step Ahead Forecasting:

- Introduced techniques for extending forecasting horizons using sliding window methods and lagged features.
- Discussed challenges of error propagation in multi-step predictions and strategies to mitigate it using ensemble and ANN-based approaches.
- Worked with real datasets, applying meta-ensemble methods to improve forecasting performance.

2 Introduction to Artificial Neural Networks (ANNs):

Participants were introduced to the structure and functionality of MLPs, including:

- Input layer: Represents the feature vectors of the dataset.
- Hidden layers: Facilitates learning of non-linear patterns.
- Output layer: Generates predictions.
- The concept of "deep networks" (with more than two hidden layers) was explained to highlight their suitability for complex problems.

3 Feature Selection and Dimensionality Reduction

Discussed techniques such as:

- Automatic feature selection methods integrated into models like LightGBM.

- Dimensionality reduction techniques such as Principal Component Analysis (PCA).

3.1.2. Follow-up discussions and progress tracking

Each seminar concluded with an interactive Q&A session and follow-up discussions to address specific challenges encountered by participants. These discussions aimed to address specific challenges encountered by participants, provide feedback on the application of discussed concepts, and ensure practical learning outcomes were achieved. Progress tracking and feedback mechanisms were implemented to ensure practical application of the concepts.

Key activities included:

1. Interactive Q&A:

- Training sessions concluded with detailed Q&A segments to allow participants to clarify issues and explore solutions collaboratively.

2. Progress Tracking:

- Feedback mechanisms were implemented to monitor participants' ability to integrate IoT and digital twin technologies into the energy performance assessment processes.

3. Documentation of Challenges:

- Challenges, such as data normalization, handling missing values, and feature selection in ML models, were identified and addressed during these discussions.

4. Actionable Follow-ups:

- Suggestions for further improvements, such as exploring new feature selection methods and dimensionality reduction techniques (e.g., PCA), were proposed for subsequent sessions.

4. Joint research activities and achievements

4.1. Preparatory phase

4.1.1. Planning, Tools and Resources for collaboration

Effective collaboration among project partners was essential to the success of the joint research activities within the SmartWins project. These activities focused on establishing clear communication channels, aligning research objectives, and leveraging appropriate tools to facilitate efficient collaboration and knowledge sharing.

1. **Kick-Off Meetings:**

- Initial meetings were held to align objectives, clarify roles, and define project deliverables.
- These sessions fostered a shared understanding of the project scope and created a roadmap for subsequent activities.

2. **Defining Milestones and Communication Frequency:**

- Project milestones were jointly developed, ensuring alignment with the project timeline and deliverables.
- Weekly or bi-weekly virtual meetings were scheduled to track progress, address issues, and set priorities for upcoming tasks.

3. **Meeting Agendas:**

- Meetings typically covered the following topics:
 - Updates on ongoing activities and tasks.
 - Discussion of challenges and resolutions.
 - Planning for upcoming phases and deliverables.

4. **Tools for Communication and Collaboration:**

- **Video Conferencing Platforms:** Tools like Microsoft Teams and Zoom were used for virtual meetings, enabling real-time discussions and decision-making.

- **Document Sharing and Management:** Cloud-based platforms such as SharePoint and Google Drive facilitated document sharing, version control, and collaborative editing.

Table 1. Overview of planned activity

Overview of planned activity			
Parts	Dates	Project Month	Description
Part 1	Oct-23	M13	Schedule a teleconference in early October and conduct a seminar focused on machine learning forecasting techniques and data preprocessing for the KTU team, emphasizing hands-on training.
Part 2	Oct-23	M13	<p>Assign tasks to the KTU team involving their historical time series datasets. This phase will encompass:</p> <ul style="list-style-type: none"> • Conducting exploratory data analysis; • Applying preprocessing techniques; • Implementing feature engineering and selection; • Developing experimental machine learning models; and • Generating and utilizing evaluation metrics. <p>The initial focus will be on fundamental machine learning algorithms.</p>
Part 3	Jan-24	M16	<p>Collect the KTU team's experimental modeling results, along with their feedback and any identified issues.</p> <p>Subsequently, CERTH will organize a teleconference to present advanced forecasting techniques—including ensembling approaches and deep learning models—as well as methods for data normalization and automated feature selection.</p>

Part 4	1/2024 – 3/2024	M16-19	<p>Conduct experiments applying the techniques learned.</p> <p>Organize a teleconference to review the KTU team's experimental results, compare these with their existing grey-box modeling approach, and discuss potential applications within KTU projects.</p> <p>This phase also involves presenting and thoroughly analyzing the available data.</p>
Part 5	3/2024-4/2024	M20	<p>Obtain comprehensive feedback from the KTU team regarding the experiments and their outcomes. Evaluate applications on data within KTU projects and conceptualize a research direction or theme.</p> <p>Follow up seminar: Principles of Effective Data Visualization</p>
Part 6	5/2024-6/2024	M20-25	<p>Initiate collaboration on a joint publication of the results. Engage with the KTU team to discuss overall outcomes, performance metrics, and encountered challenges.</p> <p>Finalize findings of the overall training exercise and research experiments</p>
Part 7	09/2024 – 10/2025	M25	<p>Collaboration on joint publication. Start drafting research publication</p>
Part 8	11/2024 – 3/2025	M25-M30	<p>Finalize publication (incl. amendments), to a peer review journal</p> <p>KTU & CERTH to collaboratively work on the final D1.3 report</p>

4.1.2. Identification of relevant research topics

The identification of research topics aimed to align the SmartWins project with its goal of advancing IoT and digital twin technologies for improving energy efficiency and resilience in smart buildings. This phase focused on identifying knowledge gaps, emerging technologies, and real-world challenges that influence the use of IoT-driven monitoring and AI-based forecasting in the built environment.

- **Literature Review:** An extensive review of academic literature, was conducted to understand state-of-the-art methodologies in digital twins, building energy performance assessment, and IoT-enabled analytics. The review helped identify critical technological bottlenecks such as sensor integration, interoperability, real-time modeling, and energy behavior under unpredictable conditions (e.g., lockdowns).
- **Focus Areas:**

Table 2. Overview of focus areas derived from literature

Focus Areas Identified in Literature Review	
Focus Area	Description
Operational Energy Modeling	Real-time monitoring, energy demand forecasting, and cross-platform integration via openBIM.
ML for Smart Systems	Use of algorithms (e.g., RF, LGBM) for forecasting, anomaly detection, and DT integration.
Crisis-Aware Modeling	Use of external data (e.g., mobility, weather) to predict shifts in consumption behavior.

1. **Operational Energy Modeling:** A primary research focus was on leveraging IoT-enabled sensing infrastructures for real-time monitoring and performance tracking of energy consumption in buildings. The use of smart meters, thermal flow sensors, and weather data provided a foundational layer for high-resolution operational insight. Additionally, data interoperability and semantic modeling frameworks—such as openBIM and linked data models—were explored to support the integration of real-time sensor data with digital twins. These frameworks facilitate seamless data exchange across heterogeneous building systems, enabling more cohesive decision support for energy optimization.
2. **Machine Learning for Smart Systems:** Another key focus was the application of machine learning (ML) algorithms to detect anomalies, forecast energy usage, and adapt control strategies dynamically within

digital twin frameworks. Experiments were also designed to integrate predictive ML outputs into digital twin control loops, enabling proactive system responses to forecasted peaks, anomalies, or crisis-triggered demand shifts. This concept is particularly aligned with the broader goals of SmartWins, which envision smart building systems as adaptive, learning-driven environments supported by digital intelligence

- **Final publication titles:**

- **Publication 1:** *Urban Traffic Congestion Prediction: A Multi-Step Approach Utilizing Sensor Data and Weather Information.*

This theme aimed to demonstrate how IoT-based traffic flow data, enriched with environmental parameters, can be used to model congestion behavior in urban environments. It validated hybrid ensemble learning for 24-hour predictive modeling of traffic, with relevance to smart city digital twin applications. The research directly supports SmartWins' goals of integrating IoT data into predictive frameworks that can power digital twin environments in the built environment.

- **Publication 2:** *Efficiency in building energy use: Pattern discovery and crisis identification in hot-water consumption data.*

This research focused on analyzing domestic hot water (DHW) energy usage in residential buildings under crisis conditions, with particular emphasis on behavioral shifts during COVID-19 lockdowns. A hybrid ensemble stacking classifier was developed to predict consumption patterns based on mobility-derived severity levels. The study provided actionable insights for adaptive control, demonstrating how behavioral data can enhance the responsiveness and energy efficiency of building systems during disruptive events.

4.2. Experimental design and implementation

The experimental design focused on developing and validating AI-based frameworks for energy performance forecasting and behavior analysis in smart buildings and cities. This work capitalized on real-world IoT datasets, ML pipelines, and simulation-based scenario modeling to explore adaptive control strategies under both standard and crisis conditions.

- **Data preparation and feature Engineering:** Data from smart meters, distributed sensors, and weather APIs formed the backbone of these experiments. Comprehensive feature engineering techniques were applied to enhance model learning and interpretability. These included:
 - Dimensionality reduction (e.g., PCA) to simplify datasets while preserving essential variance.
 - Clustering algorithms (e.g., k-means) to reveal behavioral segments or temporal usage clusters.
 - Feature selection strategies, such as recursive elimination, to isolate the most informative predictors.
 - Time series augmentation, through lagged features and temporal encoding, to embed sequence dynamics into the models.
- **Modeling Approaches:** A combination of machine learning models was employed to capture relationships between input features and key performance outcomes. These included:
 - Ensemble learning methods (e.g., gradient boosting, random forests), chosen for their robustness in handling non-linearities and feature interactions.
 - Temporal sequence models (e.g., recurrent neural networks), suited for time-series forecasting tasks involving delayed dependencies and autocorrelation.
- **Scenario-Based Simulation:** A scenario-driven methodology was used to evaluate how models respond under a variety of operating conditions. Simulations covered:
 - Normal operational states, reflecting standard daily and seasonal usage trends.
 - Crisis or disruption scenarios, modeled using external behavioral or environmental triggers (e.g., mobility restrictions, unusual weather events).
 - Short- and long-term prediction intervals, allowing the evaluation of model performance across immediate and extended time horizons.

4.2.1. Publication 1

This work aimed to develop an adaptive forecasting framework to predict multi-step traffic flow and provided insights for dynamic urban traffic management. The case study is the city of Trondheim in Norway, which has an established network of on-road sensor infrastructure measuring traffic flow. The research employed a direct multi-step forecasting approach to predict 24-hour traffic congestion.

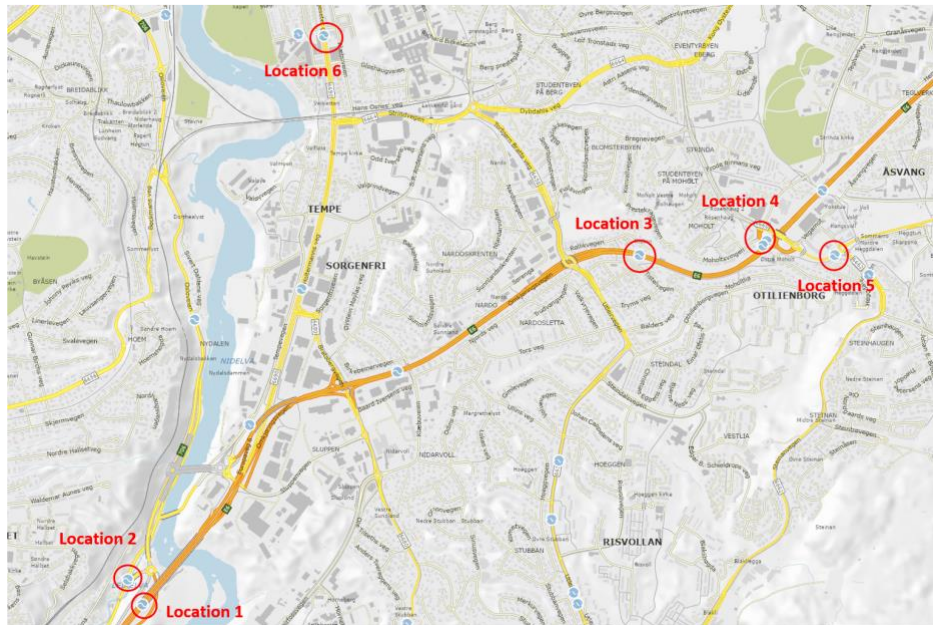


Figure 9. Locations of measured traffic flow data in the city of Trondheim.

Methodological steps:

1. Data preprocessing:

a) Data Sources:

- Traffic Flow Data: Hourly vehicle count data were collected from six strategically placed IoT traffic monitoring sites in Trondheim, Norway.
- Weather Data: External environmental variables—temperature, humidity, precipitation, and wind speed—were sourced from local meteorological stations.
- Temporal Coverage: The dataset encompassed three full years (January 2018 – December 2020), providing seasonal variability and long-term behavioral trends

b) Feature Engineering:

- **Temporal Features:** Extracted variables such as hour-of-day, weekday/weekend indicators, and seasonality to model recurrent traffic cycles.
- **Lag-Based Features:** A sliding window technique was applied to generate lagged input sequences representing prior 24-hour traffic conditions, capturing autocorrelation for sequential learning.
- **Feature Selection:** A combined approach using Recursive Feature Elimination with Cross-Validation (RFECV) and Sequential Forward Selection (SFS) was implemented to identify the most relevant predictors for the forecasting task. RFECV iteratively eliminated less informative features using a LightGBM (LGBM) estimator and cross-validation scoring to determine the optimal subset. SFS complemented this by progressively adding features that maximized forecasting accuracy (Figure 10)

Feature	Description	Unit	Selected
temp	Ambient temperature	°C	✓
feelslike	Human perceived temperature	°C	✓
dew	Dew Point	°C	
humidity	Relative Humidity	%	✓
precip	Precipitation	mm	✓
precipprob	Precipitation chance	%	
snowdepth	Depth of snow	%	✓
winddir	Direction of winds	degrees	
windspeed	Speed of wind	kph	✓
sealevelpressure	Sea level pressure	mb	✓
cloudcover	Cloud coverage	%	
visibility	Visibility	km	✓
solarradiation	Solar radiation	W/m^2	✓
solarenergy	Solar Energy	MJ/m^2	
UVindex	Intensity of ultraviolet radiation	-	

Figure 10 Weather data feature selection

c) Traffic congestion patterns

- Traffic congestion patterns were defined based on key congestion periods:

- Morning Peak: Around 8 AM during weekday commutes.
- Evening Peak: Around 4 PM, tapering off by 7 PM.
- Weekend Trends: Exhibited lower congestion overall

2. Predictive Modeling Strategy:

A Direct Multi-Step Forecasting approach was employed, generating 24 consecutive hourly predictions rather than recursive one-step forecasting. This avoided error accumulation and better captured long-range dependencies. Models developed:

- Ensemble Tree-Based Models (ETB):
 - Light Gradient Boosting Machine (LGBM).
 - Histogram-Based Gradient Boosted Regressor (HGBR).
 - Random Forest (RF) and Extra Trees (ET).
- Deep Learning Models
 - Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs).

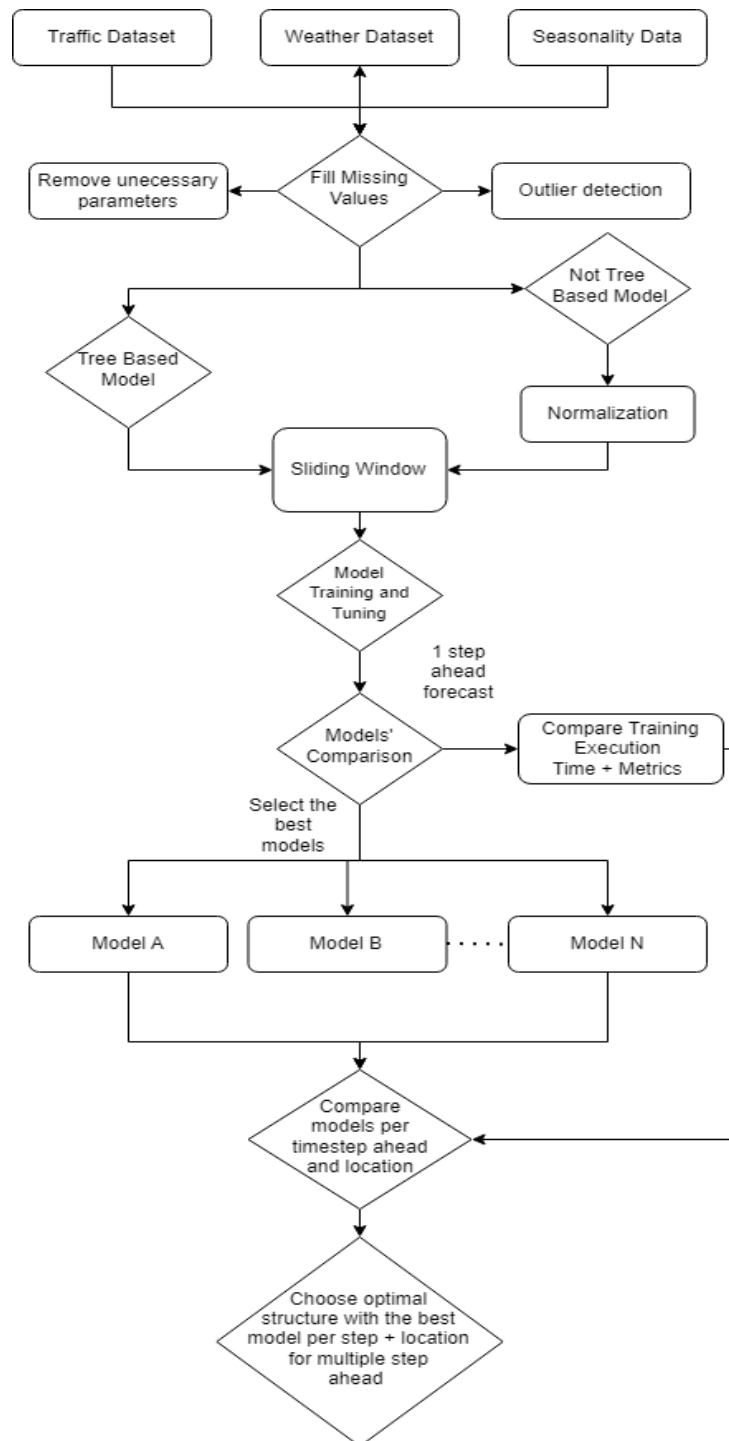


Figure 11. Overview of forecasting approach.

4.2.2. Publication 2

This study analyzed hot-water energy consumption in residential buildings during crisis periods. The objective was to capture how energy demand patterns shift in response to external behavioral disruptions (e.g., lockdowns).

In this study 10 residential apartment buildings located in Kaunas (Lithuania) were investigated (Figure 12). The investigated buildings were built between 1960s and 1980s. During the years some of the buildings were renovated while improving the thermal resistance of the building shell and installing autonomous room temperature control. The buildings are heated by district heating system (DHS) operated by local thermal energy provider. The heating supply for the buildings is regulated by the outdoor air temperature sensor. The cold water system is operated by similar principal, i.e. supplied via central district system. Hot water is prepared within building's heating unit, heated through a dedicated heat exchanger, and distributed throughout the building's hot water network for consumption. The recirculation loops ensure timely hot water supply at different parts of the building.



Figure 12. Arrangement of case study buildings in Kaunas, Lithuania.

Methodological steps:

1. Data Preprocessing, Crisis Characterization and Feature Engineering:

a) Data Sources:

- **Energy Data:** The study utilized high-resolution energy consumption data related to domestic hot water (DHW) preparation, collected from smart meters installed in 10 multi-family residential buildings in Kaunas, Lithuania. These smart meters recorded hourly thermal energy consumption, as well as flow rate and temperature measurements, across a decade-long period (2011–2021), resulting in over 480,000 entries.
- **Mobility Data (Figure 13):** Google Community Mobility Reports¹ were employed to incorporate behavioral context during the COVID-19 pandemic. The anonymized data covered the period from February 2020 to December 2021 and included two primary location categories in Kaunas:
 - Retail & Recreation (e.g., restaurants, shopping centers, cultural venues)
 - Temporal Coverage: Combined datasets spanned both pre-pandemic and pandemic periods, allowing the modeling of crisis-induced behavioral transitions.
- **Merged Dataset:** A unified dataset was constructed, combining hourly DHW consumption with severity. DHW consumption was converted into volumetric terms (m³) to allow multi-modal interpretation of energy usage trends.
- **Temporal Encoding:** Standard cyclical features (e.g., hour, weekday) were encoded to capture temporal structure in consumption behavior
- **Severity Labels:** A five-level severity scale (Baseline, Severity 1–5) was developed based on the magnitude of mobility reduction. These labels were used as input features and classification targets.

¹ Google, COVID-19 Community Mobility Report, <https://www.google.com/covid19/mobility/>. Accessed: 2025-03-019.



Figure 13. Mobility changes in relation to crisis severity levels.

2. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) was central to understanding the relationship between domestic hot water (DHW) consumption and behavioral changes across different crisis severity levels. A multi-stage data exploration pipeline was implemented to extract, segment, and visualize consumption patterns under both normal and disrupted conditions.

To analyze daily DHW consumption profiles across five severity levels and the baseline condition, the dataset was divided into six subsets. For each subset, mean hourly consumption and standard deviation (STD) values were calculated to capture typical usage patterns and intra-group variability. To ensure consistency, weekends and holidays were excluded, and each subset included 24 hourly values representing a typical weekday. Due to uneven data availability across periods, the number of days per subset varied, but the approach ensured a balanced and representative analysis of daily consumption behavior.

- **Principal Component Analysis (PCA)** was applied to reduce dimensionality and highlight dominant consumption trends. Across all severity levels, three to five components captured over 91% of the data variance.
- **K-Means Clustering**, applied to PCA-transformed data, grouped hourly consumption into four behavioral clusters: low-demand night hours, morning peaks, midday usage, and evening activity. The clustering analysis incorporated

both mean and STD values across all buildings, ensuring that differences in hot water usage among buildings were taken into account.

- **Visual analytics** techniques such as heatmaps and time-of-day density plots were generated to highlight intra- and inter-cluster consumption patterns. These visualizations effectively illustrated how usage behaviors evolved under different severity scenarios, supporting the development of forecasting and control strategies tailored to crisis-aware energy management.

3. Predictive Modeling:

A two-layer **Ensemble Stacking Classifier (ESC)** was developed to predict energy consumption behaviors under varying levels of crisis severity. The ESC structure was designed as follows:

- **First Layer – Base Learners:**
 - **LightGBMClassifier:** Utilized for its gradient-boosted decision tree efficiency on structured data.
 - **HistGradientBoostingClassifier:** A histogram-based model optimized for large datasets.
 -
- **Second Layer – Meta-Learner:**
 - **XGBoostClassifier:** Aggregated predictions from base learners and fine-tuned overall performance.

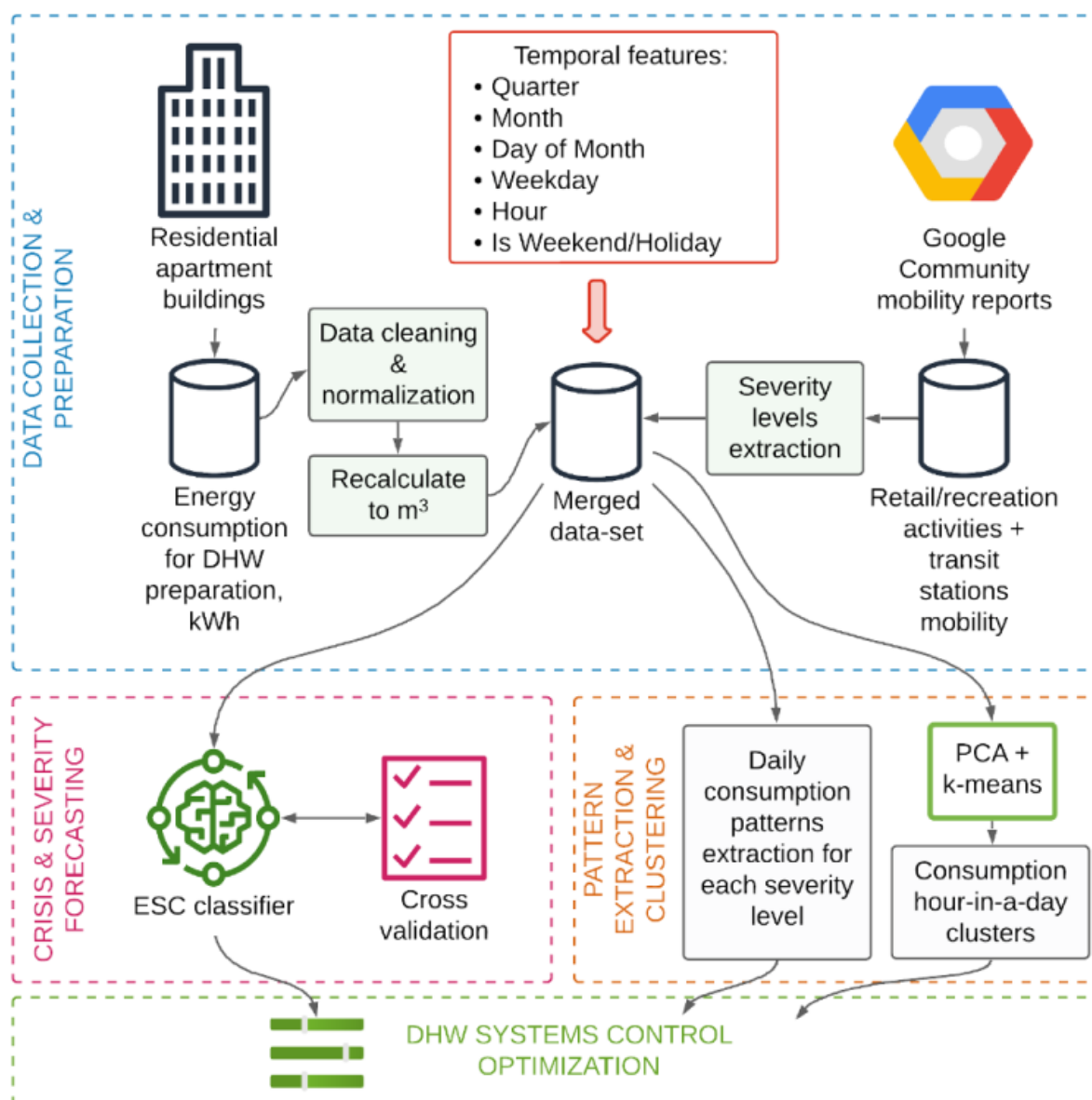


Figure 14. Overall methodological approach

4.3. Results Analysis and Insights

4.3.1. Publication 1

Insights into Congestion Patterns

The study provided a detailed analysis of urban traffic behavior, identifying temporal and contextual trends critical for transportation planning and smart city operations:

- Weekday patterns (Figure 15) exhibited two well-defined congestion peaks:
 - Morning peak around 08:00 AM, corresponding to commuting hours.
 - Evening peak between 16:00 and 19:00, gradually tapering off into the night.
- Weekend trends revealed significantly lower overall congestion, with flatter traffic curves and delayed peak times, consistent with reduced work-related travel.
- The integration of environmental variables helped capture additional variation in traffic flow, particularly during adverse weather conditions (e.g., increased congestion during rain or snowfall).

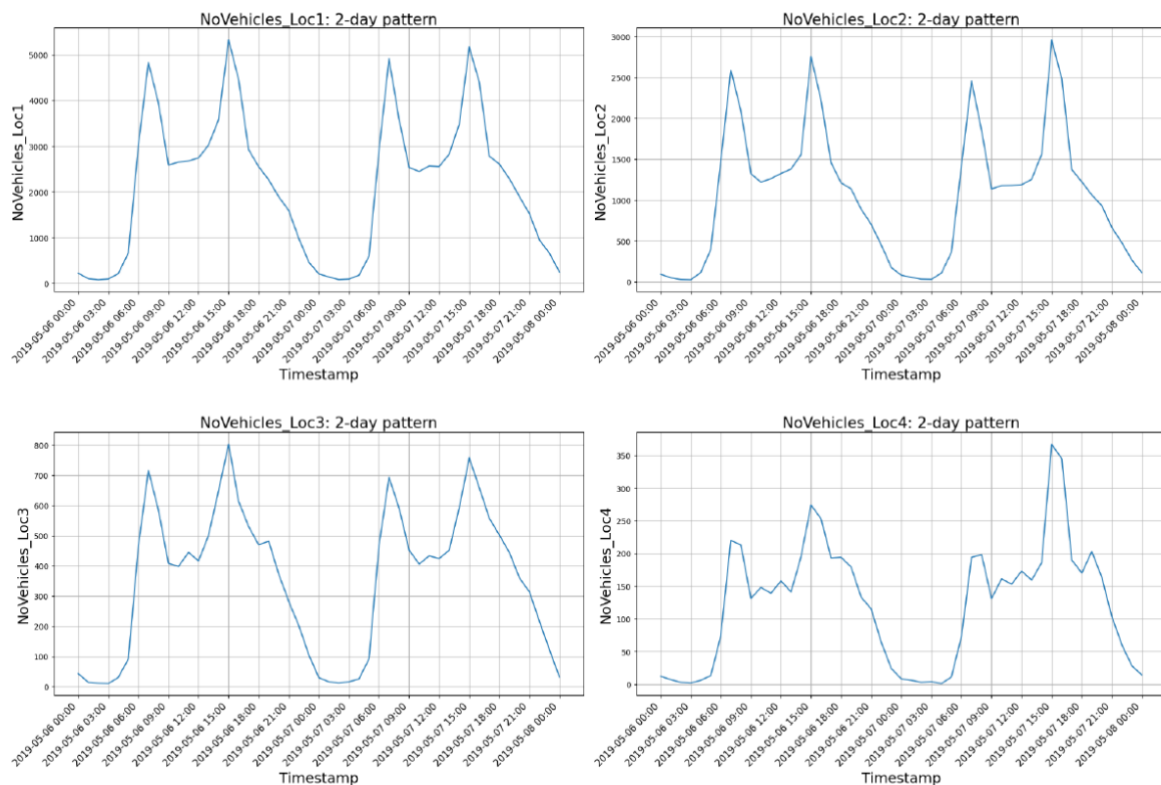


Figure 15. 48h vehicle count pattern for selected traffic locations.

Predictive Modeling Performance

The predictive modeling framework developed in this study was centered on a hybrid ensemble learning approach for multi-step traffic congestion forecasting. It integrated historical traffic and weather data from IoT-enabled monitoring stations in Trondheim to predict congestion levels across a full 24-hour horizon.

The Direct multi-step forecasting strategy was particularly effective, as it avoided the cumulative error common in recursive approaches and provided accurate hour-by-hour forecasts.

- Among the tested models, Ensemble Tree-Based Models (ETBs)—particularly LightGBM and Histogram-based Gradient Boosting Regressor (HGBR)—delivered the most consistent and accurate results across all forecasting steps.
- These models outperformed Recurrent Neural Networks (RNNs), including LSTM and GRU, especially for **longer forecasting horizons (20–24 hours)**, where tree-based models demonstrated better generalization and lower error.
- The integration of **weather data as exogenous inputs** further improved model performance by contextualizing congestion variability due to environmental factors.

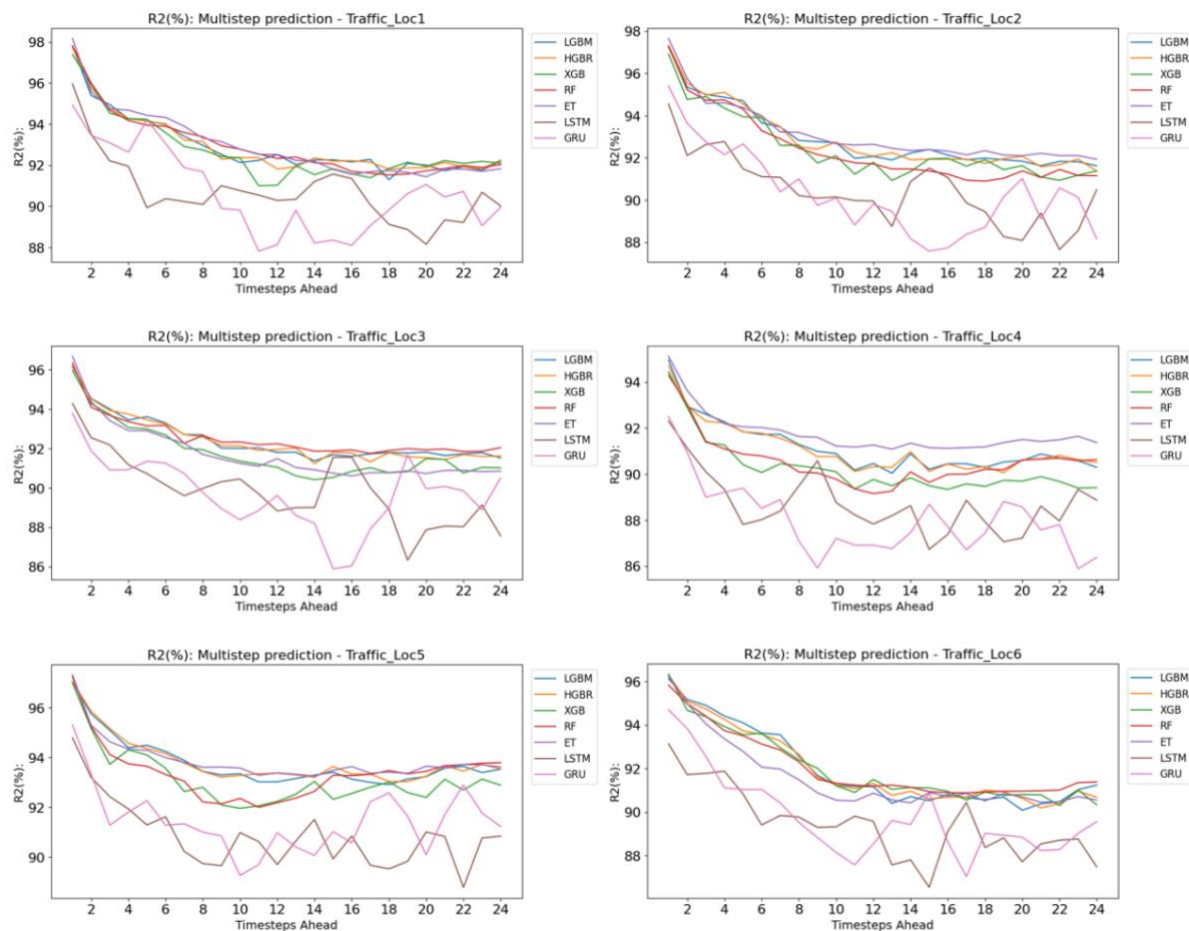


Figure 16. Direct multi-step forecasting (R^2 comparison).

4.3.2. Publication 2

Insights into Consumption Patterns

The analysis of consumption patterns across crisis phases revealed clear behavioral shifts driven by COVID restrictions:

- Morning consumption peaks, typically observed around 7 AM under normal conditions, shifted to between 12 PM and 3 PM during high-severity periods (Severity 4 and 5). This delay aligns with changes in daily routines, such as remote work and staggered household activities.
- Evening energy usage showed a significant decline during strict lockdowns. Unlike pre-crisis patterns, which exhibited a secondary evening peak, consumption flattened and aligned more closely with baseline nighttime levels, indicating reduced household activity later in the day.

- Midday demand became more prominent, particularly in Severity 3–4 phases, reflecting the convergence of remote work, homeschooling, and stay-at-home mandates.

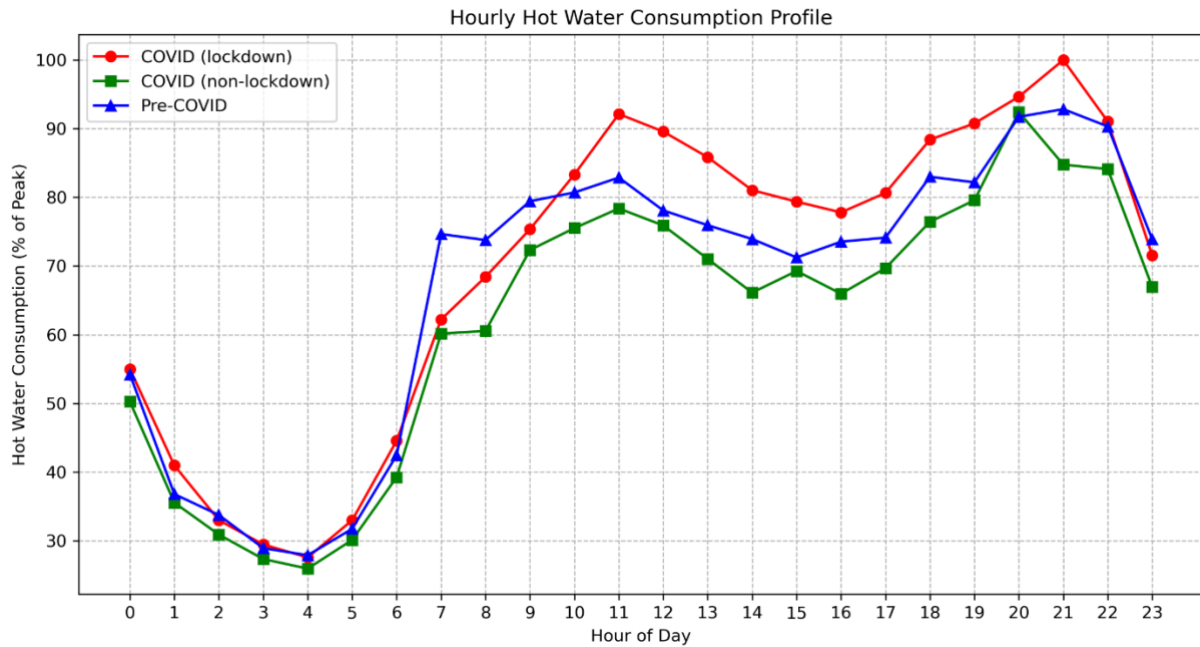


Figure 17. DHW consumption daily profiles during different periods.

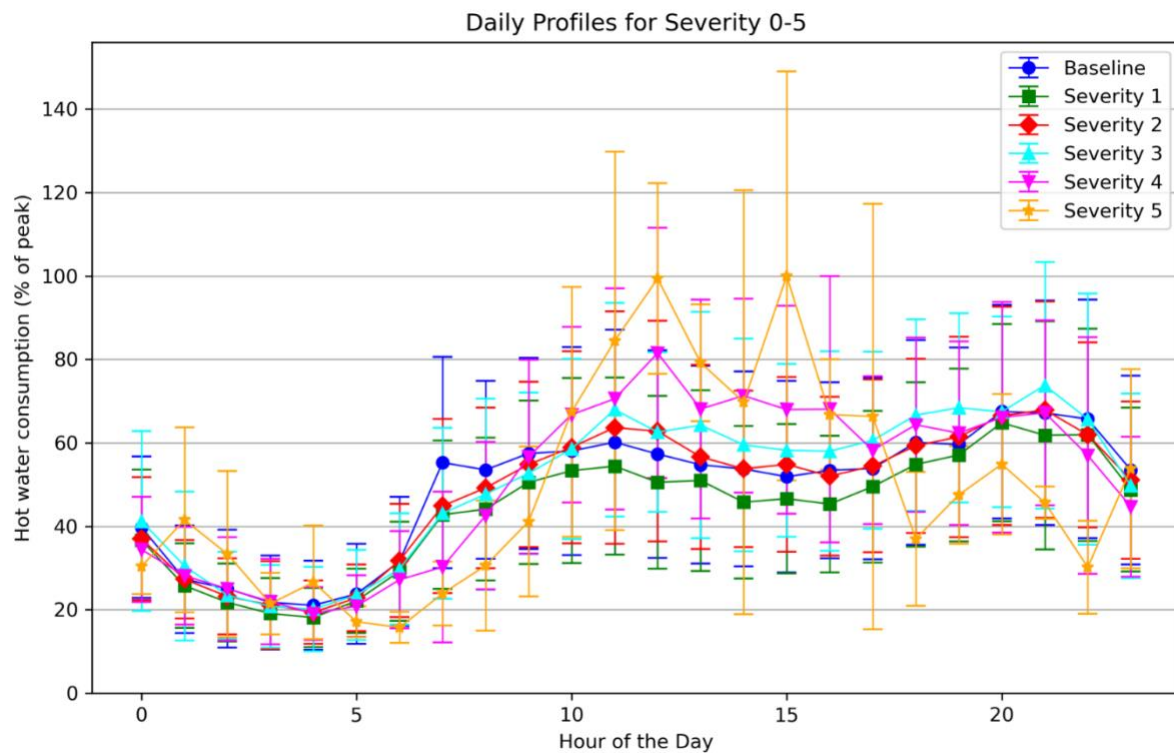


Figure 18. Extracted daily profiles based on crisis severity level.

These behavioral changes were confirmed through clustering and time-series visualization, which revealed a shift from distinct peak-and-valley profiles (in the Baseline and Severity 1–2 conditions) toward flattened, homogeneous patterns under higher severity levels.

Predictive Modeling Performance

The predictive framework, based on a hybrid ensemble stacking classifier (ESC), achieved exceptional performance in forecasting hourly DHW energy consumption patterns under varying crisis conditions. By combining multiple gradient boosting models with a meta-learner, the ESC was able to capture temporal variability in the data.

- The ESC achieved an overall classification accuracy of 99%, substantially outperforming baseline models such as LightGBM (92%).

- Evaluation metrics were consistently strong across all severity levels, with precision, recall, and F1-scores exceeding 0.99, indicating both high accuracy and balanced prediction quality.
- The ensemble approach proved particularly effective in handling intermediate disruption levels (e.g., Severity 3), where behavior patterns were less predictable.

4.1. Implications and final publication

Publication 1: Implications for Smart City Systems

The predictive framework developed in this study offers a versatile toolset for integration into smart city ecosystems, particularly when aligned with DT models of urban infrastructure. By leveraging real-time data from IoT networks and advanced forecasting techniques, this approach enables cities to anticipate and respond to mobility patterns effectively:

- **Real-Time Traffic Management:** Real-Time Traffic Management: Forecast outputs can inform adaptive traffic control systems, enabling real-time adjustments to signal timing, dynamic rerouting strategies, and congestion mitigation—especially in high-density urban corridors
- **Digital Twin Integration:** The forecasting models can be embedded into city-scale digital twin platforms, providing a virtual environment to simulate, visualize, and optimize traffic conditions under various real-world and hypothetical scenarios. This enables more informed decision-making and facilitates cross-sector coordination among transportation, energy, and environmental systems.
- **Urban Planning and Infrastructure Resilience:** The ability to forecast mobility trends provides strategic input for city planners and transport authorities, allowing them to design more responsive and resilient infrastructure. This is especially valuable in preparing for disruptions such as extreme weather events or behavioral shifts,

Publication 2: Implications for Building Smart Energy Control

The findings of this study support the development of intelligent, adaptive control strategies in smart buildings through the integration of behavioral forecasting and real-time data analytics. These capabilities are particularly relevant for stakeholders such

as building engineers, facility managers, real estate developers, and public utilities, who rely on data-driven decision support for planning and operational optimization.

- **Dynamic Energy Management:** Predictive models can be employed to automatically adjust heating and circulation schedules based on forecasted shifts in occupancy and demand, enhancing both energy efficiency and occupant comfort. Simulations indicated that such adaptive strategies could lead to an estimated 15–20% reduction in energy waste, supporting sustainability targets without compromising user needs.
- **Crisis-Aware Control:** In high-severity scenarios, control logic can prioritize midday comfort provision—when energy demand is most concentrated—while minimizing unnecessary usage during early morning and late evening hours.
- **Digital Twin Integration:** These models can be integrated into building-level digital twins, enabling real-time simulations of control scenarios and empowering facility managers with actionable insights.
- **Support for Renewable Integration:** As renewable energy sources (RES) such as solar and wind become more prevalent, the intermittency of supply introduces new challenges for balancing demand and generation. Smart control strategies in buildings—enabled by forecasting models—can help smooth consumption peaks

These results reinforce the importance of context-aware, data-driven control systems and their role in enhancing the resilience and sustainability of the built environment under both typical and crisis conditions.

A comprehensive review of reputable journals and conferences was conducted to identify appropriate dissemination channels for the project's scientific output. Priority was given to high-impact venues focusing on smart buildings, IoT, digital twins, and sustainable urban systems. Targeted journals and publishers included:

- MDPI's journals: Energies, Smart Cities
- Elsevier: Energy and Buildings, Sustainable Cities and Society,
- Other, relevant thematic journals and conferences in energy informatics, and AI for the built environment

Based on alignment with scope and readership, the following two publications were successfully prepared and published, after extensive rounds of peer review and revision.

Publication 1: Tsalikidis, N., Mystakidis, A., Koukaras, P., Ivaškevičius, M., Morkūnaitė, L., Ioannidis, D., Fokaides, P.A., Tjortjis, C., & Tzovaras, D. (2024).

Urban Traffic Congestion Prediction: A Multi-Step Approach Utilizing Sensor Data and Weather Information.

Smart Cities, 7(1), 233–253. <https://doi.org/10.3390/smartcities7010010>

Publication 2: Morkunaite, L., Pupeikis, D., Tsalikidis, N., Ivaskevicius, M., Manhanga, F.C., Cerneckiene, J., Spudys, P., Koukaras, P., Ioannidis, D., Papadopoulos, A., & Fokaides, P.A. (2025).

Efficiency in Building Energy Use: Pattern Discovery and Crisis Identification in Hot-Water Consumption Data.

Energy and Buildings, 336, 115579. <https://doi.org/10.1016/j.enbuild.2025.115579>

5. Overview of KTU gains from collaboration with CERTH

5.1. Technical skills

Through a comprehensive seminar series delivered by CERTH, KTU researchers developed critical technical skills that directly address the current challenges in implementing Digital Twins, managing IoT-enabled infrastructure, and advancing energy optimization in the built environment. Structured into theoretical and practical components, the program provided not only a solid foundation in data management, integration, and analytics, but also hands-on experience with real-world tools and coding practices, preparing researchers for complex, data-intensive scenarios in the field.

Round 1: Strengthening KTU Research Capacity in Data-Driven Digital Twins and Smart Built Environments

Starting the first part of CERTH seminars, the focus on data processing, transmission, storage, and interoperability is particularly valuable for developing Digital Twins of built environment that rely on continuous, reliable data flows from diverse sources. KTU researchers learned how to structure and manage data using standardized formats and taxonomies—essential for integrating various data types into a coherent digital replica of a physical building or system. These competencies are foundational for enabling predictive capabilities, real-time visualization, and automated control mechanisms in Digital Twin frameworks.

The exploration of openBIM formats and interoperability tools such as IFC, gbXML, IfcOpenShell, and BIMserver has equipped KTU researchers to address one of the biggest challenges in the built environment: ensuring seamless data exchange across platforms, devices, and stakeholders. This is vital for achieving open and collaborative work-flows in construction, renovation, and building operation phases, particularly as buildings increasingly rely on interoperable systems to reduce energy consumption and enhance occupant comfort.

Understanding IoT communication protocols and smart device integration—gained through practical exploration of the CERTH Smart House infrastructure—has direct application in monitoring building performance, occupant behavior, and environmental conditions in real time. These insights are critical for creating adaptive control strategies that can respond dynamically to changing conditions, such as fluctuating occupancy or weather events.

Closing the theoretical sessions, the seminars on Big Data infrastructure and tools (e.g., Apache Hadoop and Spark) addressed the pressing need to analyse massive, heterogeneous datasets generated by sensor networks in buildings. KTU researchers are now better prepared to design scalable architectures for processing and visualizing IoT data, supporting more advanced use cases like predictive maintenance, energy forecasting, and automated fault detection in building systems.

Round 2: Advancing Practical Competence in Data-Driven Energy Forecasting and AI for Smart Buildings

Building on the established theoretical foundation, the second round of seminars enabled KTU researchers to translate conceptual understanding into practical, hands-on skills, particularly relevant for the evolving needs of the built environment, energy management, and Digital Twin development. Through the use of Python-based tools and curated example code developed by CERTH, KTU researchers engaged directly

with work flows that mirror real-world challenges, such as forecasting energy consumption, handling large-scale time-series data, and designing robust, AI-driven control systems.

The first part of practical sessions focused on data preprocessing techniques that enhance model accuracy and efficiency. Throughout this session, KTU researchers gained experience in feature engineering, a cornerstone for any AI-driven energy model. These skills are increasingly necessary in the context of Digital Twins, where vast sensor data from buildings must be cleaned, structured, and transformed into usable formats that reflect real physical behaviour. The ability to apply techniques like scaling, normalization, and dimensionality reduction is fundamental not only for improving model performance but also for enabling semantic interoperability and real-time responsiveness—critical aspects of Digital Twin systems used in energy-aware building operation. Following the Python code prepared by CERTH partners, KTU researchers learned about feature extraction, selection, and transformation methods, including:

- **Handling missing data using imputation techniques.** The session explored methods for identifying and addressing gaps in time-series datasets, ensuring data integrity. Learning these skills supported KTU researchers to further improve the reliability of their machine learning models and prevent inaccurate predictions due to missing values.
- **Scaling and normalization for machine learning applications using StandardScaler and MinMaxScaler.** These techniques improved model performance by standardizing input features. Learning these techniques helped KTU researchers to ensure that different features contribute equally to model training, preventing biases caused by large numerical ranges.
- **Time-series feature engineering for forecasting models, including time-stamp handling, duplicate detection, and re-sampling strategies.** KTU researchers applied Python-based methods (datetime and matplotlib.dates) to pre-process time-series data effectively. Deepening the knowledge in feature engineering was especially useful for KTU researchers further working on accurate energy predictions, as time-series data must be properly formatted and structured to capture seasonal trends and temporal dependencies.
- **Dimensionality reduction techniques such as PCA (Principal Component Analysis).** KTU researchers learned how to reduce feature complexity while retaining the most informative components for predictive modelling. By

learning PCA KTU researchers can enhance model efficiency, reduce computational costs, and eliminate redundant information that may negatively impact performance.

The practical exercises involved using Python libraries such as Pandas, Matplotlib, Seaborn, and Scikit-Learn for data preprocessing and visualization. Additionally, KTU researchers explored real-world energy data, applying feature transformation techniques to prepare datasets for machine learning models.

During the second part of practical sessions, KTU researchers advanced their competence in deploying machine learning models tailored to building energy forecasting. Given the global shift towards energy-efficient and low-emission infrastructure, these models are central to supporting Model Predictive Control (MPC), demand-response strategies, and grid-interactive efficient buildings. The ability to train and evaluate multiple regression algorithms, including ensemble models like XGBoost, CatBoost, and LightGBM, empowers KTU researchers to compare performance and choose the most reliable predictive framework for varying use cases within energy systems and smart city contexts. As energy forecasting plays a critical role in building management, grid optimization, and sustainability efforts, which are relevant research areas for KTU, this session provided insights into advanced machine learning techniques tailored for energy applications. In particular participants explored:

- **Regression models for energy demand prediction**, including: XGBoost; Random Forest; K-Nearest Neighbors (KNeighborsRegressor); LightGBM; CatBoost; AdaBoost; Bagging and Voting Regressors. Learning these regression models equipped KTU researchers with the ability to develop highly accurate energy demand prediction models, essential for optimizing building management, grid operations, and sustainability strategies. By understanding and applying diverse algorithms like boosting, bagging, and ensemble methods, KTU researchers learned to enhance model robustness, improve forecasting reliability, and support data-driven decision-making in smart energy systems and Digital Twin applications.
- **Time-series forecasting techniques**, including: ARIMA and Prophet models for long-term energy predictions; Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks for deep learning-based forecasting; Hyperparameter tuning and model optimization, utilizing GridSearchCV and RandomizedSearchCV to improve model performance as well as model evaluation metrics, such as mean squared error, R2 score, mean absolute error, and mean absolute percentage error. Learning these time-series forecasting

techniques enabled KTU researchers to develop more precise and adaptive energy consumption predictions, improving building energy management, smart grid operations, and sustainability planning by leveraging both statistical models and deep learning approaches.

The practical exercises in this session focused on training and evaluating machine learning models for energy forecasting, incorporating feature engineering, model selection, and hyper-parameter tuning. KTU researchers applied their knowledge to real-world energy datasets, optimizing predictions for building energy management and smart grid applications.

The final practical seminar delivered by CERTH introduced deep learning architectures and ensemble forecasting techniques, enabling KTU researchers to leverage neural networks and advanced regression models for improved energy forecasting accuracy. The focus on deep learning architectures prepared KTU researchers to a higher level of modelling complexity. Learning tools like Artificial Neural Networks (ANNs), LSTM, GRU, and Conv1D networks reflects a key step forward in handling sequential data and adaptive forecasting, essential for buildings equipped with real-time sensors and IoT infrastructure. These methods are particularly relevant to future challenges in smart building automation, fault detection, and proactive energy system control. Understanding model evaluation metrics and optimization techniques also positions researchers to engage in rigorous validation and deployment of models within operational settings. Specific techniques learned were:

- **Artificial Neural Networks (ANNs) with TensorFlow/Keras**, where KTU researchers learned how to structure and optimize deep learning architectures for energy forecasting.
- **Recurrent Neural Networks (RNNs)**, specifically Long Short-Term Memory (LSTM), Bidirectional LSTMs, and Gated Recurrent Units (GRU), which specialize in handling sequential energy data for time-series forecasting.
- **Convolutional Neural Networks (Conv1D)** for feature extraction in time-series analysis, enhancing model performance.
- **Training and evaluation of neural networks**, utilizing Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Accuracy as key performance indicators.
- **Drop-out layers** for preventing over-fitting and **Time-distributed layers** for sequence prediction.

The hands-on session allowed KTU researchers to train and evaluate different deep learning models, gaining practical experience in building AI-driven forecasting systems for energy applications.

Together, these practical exercises equipped KTU researchers with a full-cycle understanding of how to move from raw sensor data to actionable insights and control strategies, a crucial capability for realizing the promise of data-centric, AI-augmented Digital Twins. As the building sector moves toward increased automation and intelligence, these skills are not only academically relevant but instrumental in addressing real-world challenges related to climate adaptation, energy efficiency, and the digital transformation of the built environment.

Throughout the theoretical and practical sessions delivered by CERTH, KTU researchers significantly enhanced their expertise in data management, machine learning, and predictive analytics. The combination of theoretical foundations and hands-on coding sessions enabled them to develop advanced skills in Digital Twin integration, energy forecasting, and deep learning models. These newly acquired competencies will support future research in smart energy management, sustainable building solutions, and AI-driven predictive modelling, strengthening KTU's role in cutting-edge technological advancements.

5.2. Machine Learning methods application

By contributing to the research titled *"Urban Traffic Congestion Prediction: A Multi-Step Approach Utilizing Sensor Data and Weather Information"*, KTU researchers deepened their practical and theoretical understanding of machine learning-driven forecasting, particularly in spatio-temporal systems. This study provided an opportunity to apply advanced data science techniques to a real-world urban mobility challenge, fostering skills that are broadly transferable to related fields such as Digital Twins, smart cities, and energy systems modelling.

KTU researchers strengthened their data preprocessing and integration skills, working with diverse data sources, including IoT-based traffic sensors and weather data, and applying techniques like linear interpolation, feature extraction, and sliding windows for time-series expansion. These methods improved their ability to model complex temporal systems, a competency directly applicable to building energy forecasting and dynamic occupancy modelling in smart infrastructure.

In terms of modelling, the team gained valuable experience in feature selection using Recursive Feature Elimination with Cross-Validation (RFECV), and in evaluating models using metrics like R^2 and CVRMSE. They assessed and compared a wide array of models—starting with 18 and narrowing down to 6 after initial benchmarking—eventually identifying LSTM and GRU neural networks as the most effective for short-term traffic forecasting. This process enhanced their understanding of deep learning architectures suited for sequential data and introduced hyper-parameter tuning as a key optimization step.

KTU researchers learned to recognize the limitations of traditional approaches when spatial structure is ignored, and the potential for graph-based modelling, informed by space syntax theory, to address spatial dependencies in infrastructure systems. This insight is particularly relevant for expanding Digital Twin capabilities beyond static monitoring to include network-based behaviour modelling—such as water flow, occupant movement, or power distribution.

5.3. Multivariate data analysis methods application

As part of the research and publication process for the study *"Efficiency in Building Energy Use: Pattern Discovery and Crisis Identification in Hot-Water Consumption Data"*, KTU researchers gained extensive technical skills related to data processing, predictive modelling, and advanced energy analytics. The publication required expertise in machine learning, time-series analysis, and control system optimization, particularly in the context of crisis conditions and urban mobility changes. The knowledge gained from CERTH partners during the theoretical and practical seminar series enabled KTU researchers to perform high-quality analysis that was later published in a key journal for civil engineering field "Energy and Buildings". Key competences learned that supported the preparation of the manuscript included:

1. **Advanced Data Processing & Feature Engineering techniques.** KTU researchers learned how to handle and pre-process large-scale datasets, ensuring data quality for further analysis. They developed skills in data cleaning, filtering, and normalization, allowing them to remove inconsistencies and standardize consumption patterns. Additionally, they gained experience in temporal feature extraction, identifying key time-dependent patterns such as hour-of-day trends, seasonal variations, and holiday effects on energy consumption. By integrating external mobility datasets, they learned how to correlate human activity trends with

resource usage, a crucial skill for data-driven decision-making in smart energy systems.

2. **Predictive Modelling for Crisis Identification.** Through the development of a Hybrid Ensembling Stacking Classifier (ESC), KTU researchers enhanced their understanding of ensemble learning techniques and their advantages in predictive modelling. They learned how to fine-tune hyper-parameters to improve model performance, ensuring high accuracy in crisis severity prediction. By applying 5-fold cross-validation, they gained expertise in model validation and generalizability assessment, which are essential for deploying robust machine learning models in real-world applications. Their experience in anomaly detection and forecasting techniques provided them with critical skills for analysing unexpected consumption patterns in energy systems.
3. **Time-Series Forecasting & Pattern Extraction.** KTU researchers developed a deeper understanding of time-series analysis by working with Principal Component Analysis (PCA) and clustering techniques. They learned how to reduce data dimensionality while preserving critical information, improving computational efficiency and interpretability. Through k-means clustering, they gained practical knowledge of how to classify daily DHW usage patterns into distinct behavioural groups. They also enhanced their ability to identify shifts in consumption trends due to external factors, such as crisis events, refining their expertise in pattern recognition and trend analysis for smart building control.
4. **Control Optimization for Energy Management.** By modelling hot water system components, KTU researchers gained knowledge of energy flow dynamics in real-world systems. They learned how to develop automated control strategies based on ISO 52120-1:2021, applying industry standards to energy optimization. Additionally, they acquired skills in predictive control for district heating and energy storage, allowing them to design more resilient and adaptive energy management systems. Their work with AI-driven decision-making models further reinforced their understanding of automated energy efficiency strategies in smart buildings.

Through this publication process, KTU researchers applied the gained technical expertise in advanced data analytics, machine learning, and real-time energy optimization. The integration of mobility data with predictive modelling allowed them to develop highly adaptive DHW control strategies, improving energy efficiency and resilience in smart buildings. These newly acquired technical skills will directly support

future research in Digital Twins, AI-driven energy management, and crisis-resilient infrastructure planning. Additionally during the publication writing process, KTU researchers have learned how to work with LaTeX language for publication preparation, and prepared both manuscripts in Overleaf environment.

6. Conclusions

The activities conducted within Task 1.3 of the SmartWins project have yielded valuable insights into the integration of Internet of Things (IoT) technologies, Digital Twins (DTs), and advanced machine learning (ML) for performance assessment and intelligent management of smart buildings. The synergy between these technologies enables continuous data acquisition, real-time monitoring, and predictive analytics—capabilities that are essential for the transition toward carbon-neutral and resilient built environments.

Experimental results from real-world case studies validated the effectiveness of hybrid ensemble and neural network-based forecasting techniques. The two scientific publications derived from these activities demonstrated that ensemble models—particularly those based on tree-based architectures and stacked learning strategies—significantly outperformed conventional approaches in both urban traffic forecasting and building energy consumption prediction under atypical (crisis) conditions. These findings highlight the transformative potential of integrating IoT, DTs, and ML not only for optimizing energy usage but also for enhancing the overall sustainability and operational efficiency of urban infrastructures.

A particularly significant outcome of this task has been the deepened collaboration between KTU and CErTH, which substantially enhanced the technical capacities of KTU researchers. Through a structured seminar series, practical case studies, and joint research efforts, KTU personnel developed a comprehensive skill set spanning data management, machine learning, time-series forecasting, and control system modelling. These competencies are vital for addressing real-world challenges in the built environment, such as improving energy efficiency, enabling intelligent control strategies, and managing uncertainty in dynamic, sensor-rich systems.

The knowledge gained through hands-on coding exercises, direct experimentation with real datasets, and advanced analytics enabled KTU researchers to progress from raw sensor data to the development of actionable decision-making tools. Their demonstrated ability to apply statistical and AI models, integrate heterogeneous data sources, and simulate control strategies positions them to take a leading role in the design and implementation of resilient, adaptive, and energy-aware digital twin environments. The successful application of these skills in peer-reviewed publications

further underscores their relevance, scientific value, and practical impact, while also contributing to the academic development and international visibility of the SEBERG group.

Finally, the structured knowledge transfer activities played a critical role in aligning project milestones and building a shared understanding of the opportunities and challenges associated with deploying smart technologies for energy management. This collaborative capacity-building process sets a strong foundation for the long-term sustainability of SmartWins outcomes and the future leadership of KTU in the digitalization of the built environment.

Disclaimer

This document contains information which is proprietary to KTU. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of KTU. The information, views and tips set out in this publication are those of KTU and the project group and cannot be considered to reflect the views of the European Commission or its support services.

References

- Agostinelli, S., Cumo, F., Guidi, G., & Tomazzoli, C. (2021). Cyber-physical systems improving building energy management: Digital twin and artificial intelligence. *Energies*, 14(8). <https://doi.org/10.3390/EN14082338>
- Baghalzadeh Shishehgarkhaneh, M., Keivani, A., Moehler, R. C., Jelodari, N., & Roshdi Laleh, S. (2022). Internet of Things (IoT), Building Information Modeling (BIM), and Digital Twin (DT) in Construction Industry: A Review, Bibliometric, and Network Analysis. *Buildings*, 12(10). <https://doi.org/10.3390/BUILDINGS12101503>
- Bashir, M. R., Gill, A. Q., & Beydoun, G. (2022). A Reference Architecture for IoT-Enabled Smart Buildings. *SN Computer Science*, 3(6). <https://doi.org/10.1007/s42979-022-01401-9>
- Bortolini, R., Rodrigues, R., Alavi, H., Vecchia, L. F. D., & Forcada, N. (2022). Digital Twins' Applications for Building Energy Efficiency: A Review. *Energies* 2022, Vol. 15, Page 7002, 15(19), 7002. <https://doi.org/10.3390/EN15197002>
- "D²EPC" Horizon project. (2021). *D²EPC Framework Architecture and Specifications*.
- "D²EPC" Horizon project. (2022). *D²EPC D3.3 - Building Digital Twin*.
- Davila Delgado, J. M., & Oyedele, L. (2021). Digital Twins for the built environment: learning from conceptual and process models in manufacturing. *Advanced Engineering Informatics*, 49. <https://doi.org/10.1016/J.AEI.2021.101332>
- Elfarri, E. M., Rasheed, A., & San, O. (2022). *Artificial intelligence-driven digital twin of a modern house demonstrated in virtual reality*. <https://arxiv.org/abs/2212.07102v2>
- Francisco, A., Mohammadi, N., & Taylor, J. E. (2020). Smart City Digital Twin–Enabled Energy Management: Toward Real-Time Urban Building Energy Benchmarking. *Journal of Management in Engineering*, 36(2). [https://doi.org/10.1061/\(asce\)me.1943-5479.0000741](https://doi.org/10.1061/(asce)me.1943-5479.0000741)
- Justas Brazauskas, Matt Danish, Vadim Safronov, Rohit Verma, Richard Mortier, & Ian Lewis. (2022). *CDBB West Cambridge Digital Twin: Lessons Learned*.
- Kaewunruen, S., Rungskunroch, P., & Welsh, J. (2019). A digital-twin evaluation of Net Zero Energy Building for existing buildings. *Sustainability (Switzerland)*, 11(1). <https://doi.org/10.3390/SU11010159>

- Moretti, N., Xie, X., Merino, J., Brazauskas, J., & Parlikad, A. K. (2020). An openBIM approach to IoT integration with incomplete as-built data. *Applied Sciences (Switzerland)*, 10(22), 1–17. <https://doi.org/10.3390/app10228287>
- Petri, I., Rezgui, Y., Ghoroghi, A., & Alzahrani, A. (2023). Digital twins for performance management in the built environment. *Journal of Industrial Information Integration*, 33. <https://doi.org/10.1016/J.JII.2023.100445>
- “Re-Cognition” Horizon Project. (2019). *Re-Cognition D1.4 - System Architecture*.
- “Re-Cognition” Horizon project. (2019). *Re-Cognition D3.1 - Common Information Model and Device Managers*.
- Shahzad, M., Shafiq, M. T., Douglas, D., & Kassem, M. (2022). Digital Twins in Built Environments: An Investigation of the Characteristics, Applications, and Challenges. *Buildings*, 12(2). <https://doi.org/10.3390/BUILDINGS12020120>
- Singh, M., Fuenmayor, E., Hinchy, E. P., Qiao, Y., Murray, N., & Devine, D. (2021). Digital twin: Origin to future. In *Applied System Innovation* (Vol. 4, Issue 2). MDPI AG. <https://doi.org/10.3390/asi4020036>
- Wang, P., & Luo, M. (2021). A digital twin-based big data virtual and real fusion learning reference framework supported by industrial internet towards smart manufacturing. *Journal of Manufacturing Systems*, 58, 16–32. <https://doi.org/10.1016/j.jmsy.2020.11.012>
- Yang, Y., Pan, Y., Zeng, F., Lin, Z., & Li, C. (2022). A gbXML Reconstruction Workflow and Tool Development to Improve the Geometric Interoperability between BIM and BEM. *Buildings*, 12(2). <https://doi.org/10.3390/buildings12020221>
- Zhao, L., Zhang, H., Wang, Q., & Wang, H. (2021). Digital-Twin-Based Evaluation of Nearly Zero-Energy Building for Existing Buildings Based on Scan-to-BIM. *Advances in Civil Engineering*, 2021. <https://doi.org/10.1155/2021/6638897>