# BTS: Building Timeseries Dataset: Empowering Large-Scale Building Analytics

**Arian Prabowo**[1], **Xiachong Lin**[1], **Imran Razzak**[1], **Hao Xue**[1], **Emily W. Yap**[2], **Matthew Amos**[3], **Flora D. Salim**[1]

[1]CSE, UNSW *, Sydney NSW 2052
[2]SBRC †, UOW, Wollongong NSW 2522
[3]Energy, CSIRO ‡, Newcastle NSW 2304
{arian.prabowo, imran.razzak, hao.xue1, flora.salim}@unsw.edu.au
dawn.lin@student.unsw.edu.au eyap@uow.edu.au matt.amos@csiro.au

## Abstract

Buildings play a crucial role in human well-being, influencing occupant comfort, health, and safety. Additionally, they contribute significantly to global energy consumption, accounting for one-third of total energy usage, and carbon emissions. Optimizing building performance presents a vital opportunity to combat climate change and promote human flourishing. However, research in building analytics has been hampered by the lack of accessible, available, and comprehensive real-world datasets on multiple building operations. In this paper, we introduce the Building TimeSeries (BTS) dataset. Our dataset covers three buildings over a three-year period, comprising more than ten thousand timeseries data points with hundreds of unique classes. Moreover, the metadata is standardized using the Brick schema. To demonstrate the utility of this dataset, we performed benchmarks on the multi-label timeseries classification task. This task represent an essential initial step in addressing challenges related to interoperability in building analytics. Access to the dataset and the code used for benchmarking are available here: https://github.com/cruiseresearchgroup/DIEF_BTS

## 1 Introduction

**Importance of building analytics.** Building analytics, also known as data-driven smart building [12], involves the automated adjustment of building operations to minimize emissions and costs, optimize energy usage, and enhance indoor environmental quality and occupant experience, including comfort, health, and safety [72]. This is particularly crucial given that buildings account for a third of global energy usage and a quarter of global carbon emissions, comparable to the transport sector [27]. Optimizing building performance has the potential to significantly mitigate climate change and promote human well-being.

**Literature gaps.** This paper addresses two critical gaps in building analytics research. Firstly, in Section 2.1, we highlight the scarcity of publicly available and freely accessible datasets on comprehensive real-world building operations, as exemplified in Table 1. While LBNL59 [49, 36] is the only dataset that captures various aspects of building operations comprehensively, it only includes data from a single building.

*School of Computer Science and Engineering (CSE), University of New South Wales (UNSW).
†Sustainable Buildings Research Centre (SBRC), University of Wollongong (UOW).
‡Commonwealth Scientific and Industrial Research Organisation (CSIRO).

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-4232

Table 1: **Comparing the scope of representative datasets for building analytics**. Only datasets on real-world building operations that are available, accessible are presented. Note that non-intrusive load monitoring (NILM) is not a single dataset but a task that usually use similar datasets. Similarly, AshraeOB is also a collection of dataset.

| Year | Dataset | Unique Class | Scope |
|---|---|---|---|
| 2013 | SLRHOME [5] | 3 | Aggregate energy load and generation |
| 2014 | LCLD [79] | 2 | Aggregate energy load and tarriff |
| 2015 | UCI [78] | 1 | Aggregate energy load |
| 2017 | BGD2 [51] | 18 | Detailed energy load |
| 2020 | LBNL59 [36, 49] | 35 | **Comprehensive** |
| 2021 | AshraeOB [18, 47] | 76 | Occupancy and their factors (e.g. lighting, indoor climate) |
| Varies | NILM [72] | Varies | Detailed energy load |
| 2024 | **BTS (Ours)** | 215 | **Comprehensive** |

This limitation underscores the need for datasets covering multiple buildings to address the second gap: interoperability in building analytical models. Interoperability is crucial for scalability, allowing models to be applied across diverse buildings with differing characteristics such as climate, usage, size, regulations, budget, and architecture. This challenge is discussed further in Section 2.2. Additionally, such datasets inherently possess properties of interest to machine learning research, such as domain shift, multimodality, imbalance, and long-tailedness, which are discussed further in Section 2.3.

**Building TimeSeries (BTS): A new dataset.** In this paper, we introduce a new anonymized building analytics dataset sourced from three buildings located in undisclosed regions across Australia. Spanning a three-year period, our dataset encompasses over ten thousand timeseries data points, featuring a diverse array of 240 unique classes. Notably, this surpasses the ontological breadth of LBNL59 by more than threefold. These ontologies serve as standardized categorizations of building timeseries data, including parameters like `Temperature_Setpoint` and `Voltage_Sensor`. The breadth of ontologies within our dataset enables researchers to explore buildings with more intricate analytics setups, facilitating deeper insights into building dynamics and performance. Furthermore, the metadata are standardized using the popular Brick schema [7], ensuring consistency and compatibility across analyses.

**A Benchmark.** To demonstrate the utility of this dataset, we conducted benchmarks on a machine learning model interoperability task: multi-label timeseries classification. One of the initial steps in achieving building analytics interoperability is to map thousands of heterogeneous timeseries generated from sensors and actuators to a standardized ontology, such as the Brick schema [7]. This is also known as the timeseries ontology classification task [67].

We also performed an additional benchmark on a zero-shot forecasting task [19, 28]. This explores scenarios where a building manager deploys a pre-trained model without fine-tuning. This task is more complex than typical setups because the model must generalize to an arbitrary number of timeseries, various permutations of their ontologies, and their relationships [45]. The details can be found in Appendix D.

**Contribution.** This paper introduces the Building TimeSeries (BTS) dataset, addressing critical gaps in publicly available building analytics datasets. Existing datasets often lack accessible, available, comprehensive, real-world, building operations data, hindering progress in building analytics research. While some datasets like LBNL59 offer a holistic view, they are limited to single buildings, impeding efforts to achieve interoperability in building analytics models. BTS fills this void by providing data from three diverse buildings, spanning a three-year period and encompassing over ten thousand timeseries data points and 240 unique classes. Morever, BTS inherently possess properties relevant to machine learning research, including domain shift, multimodality, imbalance, and long-tailedness. Furthermore, we conduct a benchmark on a machine learning model interoperability task — multi-label timeseries classification — demonstrating BTS's utility in addressing challenges related to interoperability in building analytics. Overall, BTS dataset advances the pursuit of optimizing building performance, ultimately aiding efforts to mitigate climate change and enhance human flourishing.

## 2 Related Works

### 2.1 Existing Datasets

To write this section, we reviewed of the building datasets utilized in the literature. We found that, in most cases, the datasets are private, static, simulation-based, or limited in ontology. Although our review is not systematic as this is not a review paper, our search was sufficiently extensive to ensure the validity of our findings. The datasets discussed here are primarily derived from five recent review papers [59, 72, 39, 40, 44] along with our own collections. This would have included earlier surveys such as [6]. Table 4 in the appendix list the works mentioned in this section.

**Availability and Accessibility**. Most research on building analytics uses private datasets [82]. This is due to security and privacy concerns of building owners and occupants. This is prevalent across many aspects of building analytics, from HVAC [69, 33, 77, 71, 30, 29, 20], energy use [60, 61], and more holistic systems [34, 35, 23, 42, 43, 67].

Some datasets are publicly accessible, but not for free, such as Pecan Street [14], or not freely available, such as ecobee [21]. Notably, the Mortar dataset [22], which comprises data from 90 buildings and over 9.1 billion data points, is currently unavailable due to cloud deployment issues at the time of writing.

**Building Operation**. Most public datasets such as EUBUCCO [53] only contain static information such as type, height, and construction year. However, these datasets do not contain sufficient information on building operation. Others contain more extensive information, such as PLUTO [17] and GBMI [10] with more than 70 fields and 380 fields respectively, or building polygons [87] and 3D shapes [9].

While many public datasets include time information, they are often too sparse (yearly) to be useful for building analytics, which require at least daily data. Examples include the popular CBECS [16], and larger ones like BERTOOL [75] and CENED+2 [68], each containing about a million instances.

**Real-World and Not Simulation**. Simulations, while valuable, present limitations due to their reliance on assumptions that may not accurately reflect real-world building systems and human behaviors [95, 72]. Results have been shown to diverge from actual telemetry data in multiple studies [74, 1, 76]. These simulations are often calibrated to match existing datasets such as BEM4CBECS [2, 91, 92, 90] which are based on the CBECS dataset [16], while ResStock [84] and ComStock [58] are based on data from 2.3 million meters in the US [85]. Another notable examples are CityLearn Challenge Series [81, 54, 57, 56]. Not all simulations are software-based. There are also hardware-in-the-loop laboratory setup [65, 64].

**Whole Building Scope**. The few remaining datasets are listed on Tab. 1. They have limited scope, and does not fully capture the entire building as a holistic system. For example, most datasets are focused only on aggregated energy load (UCI [78]), or disaggregated (ASHRAE [32, 31, 37], BDG [52, 51], NILM [59]), or when combined with generation [5], or price [79]. Others focuses on occupancy patterns [25, 24, 18, 47] or water [13, 70].

To our knowledge, LBNL59 [49, 36], a medium-sized office building in Berkeley, is the only comprehensive existing dataset. Our dataset complements this dataset by introducing three new buildings, with more diverse ontology. This allows the exploration various transfer learning techniques to ensure that machine learning models are interoperable between buildings. In Section 3.2, we make a detailed comparison of LBNL59 with our dataset.

### 2.2 Relevant Challenges in Building Analytics

The standardization of building timeseries data overcomes the challenge of interoperability and scalability that can give rise to greater widespread adoption of energy flexibility in a systematic manner. Achieving zero-energy buildings has two conflicting optimization goals: to maximise occupant comfort and indoor environmental quality, and to minimise carbon emissions and operating costs [41]. It involves two components: the building model that represents the thermodynamics and energy behavior of a building and its components such as its construction, materials, and HVAC system, and secondly, a control strategy to automate the control operations.

 https://doi.org/10.52202/079017-4232

Obtaining a building model involves expert knowledge and significant time to develop and validate. This is further amplified by requiring individual models for each building. These models can be white-box (physics-based) [95, 80], black-box (data-driven), or grey-box (hybrid) [50, 46]. Our dataset and benchmark experiment, which automate timeseries data classification, help address this challenge by reducing the time and cost associated with building key components of these models.

In comparison to building models, there has been a significant focus on optimising building control operations and transitioning from conventional rule-based approaches to model predictive control or data-driven methods [50]. The Building Optimization Framework or BOPTEST [11] exists to enable the development and benchmarking of building control strategies. The performance of a control strategy or algorithm is evaluated on a virtual "test case". Currently, these test cases are simulation physics-based models of ideal buildings developed on Spawn [83] (a co-simulation of Modelica and EnergyPlus) and act as emulators. In their paper, Blum et al. [11] make the contrasting argument that simulation-based test cases offer advantages over existing challenges when testing in real buildings, such as being time-consuming and subject to stochastic events.

However, accessing publicly available and anonymized building timeseries data from various non-residential building types acts as a commodity to reduce the time to develop individual hybrid building models. On one hand, using data from real buildings can be used to calibrate and interpolate lesser-known parameters, while maintaining moderate interpretability. And on the other hand, using standardized timeseries data such as the datasets introduced here aids in scalability and deployability to build generalized multi-zone environments and substituting with data from another building system or zone.

More broadly, there are various other applications of this dataset. **Generative AI for Privacy-Preserving Data Sharing**: Explore the use of generative AI to create synthetic building timeseries data, enabling building owners to contribute data for research while safeguarding sensitive information. **LLM Integration for Natural Language Interaction**: Investigate methods to integrate LLMs with building timeseries data, allowing various stakeholders such as building operators to interact with and query the data using natural language. **Redeployability**: By using a standardised ontology to describe the building, and linking timeseries data to the building model, applications (e.g. measurement and verification, chiller scheduling, occupant comfort) can be written to deploy against a fleet of buildings without a deep understanding of the building topology, such as those provided within this dataset.

### 2.3 Relevant Challenges in Machine Learning (ML) Research

**Domain shift and domain adaptation.** In the realm of ML research, one challenge is in domain adaptation, particularly about the diverse characteristics of buildings. These variations encompass factors such as climate, usage, size, regulations, budget, and architecture, resulting in notable distribution shifts. Consequently, traditional ML methodologies fall short in address these discrepancies. Therefore, the development and implementation of domain adaptation techniques [4, 3, 73, 26] are crucial to ensure model generalization across different buildings. Additionally, the usual alternative of employing large foundational models [94] is impractical because privacy and security concerns limit the availability of extensive building datasets for training. Moreover, as shown in Section 3.3.2, the unique permutation of ontologies in each building further complicates the scenario, necessitating novel approaches capable of handling arbitrary permutations effectively [45]. This is an issue since many timeseries architecture do not allow the model to input and output an arbitrary number of variate [86].

**Multimodal Learning with knowledge graphs (KG) and unbalanced multivariate timeseries (MVTS) with long tails.** While many studies focus on MVTS data in conjunction with spatial graph [62, 63], video, image, audio, and text data [15, 89], research on MVTS with knowledge graphs is scarce. Our dataset enable such research as it contains the Brick schema which is a KG on building metadata, describing relationship between the timeseries in the MVTS. Our dataset is also challenging because it is unbalanced and featuring distributions long tails. As shown in Section 3.3.2, some classes, like `Chilled Water Differential Temperature Sensor`, might only have one or two instances in the entire dataset, or, like `Alarm`, have zero values for most of the time . These challenges could fuel the developments of innovative techniques.

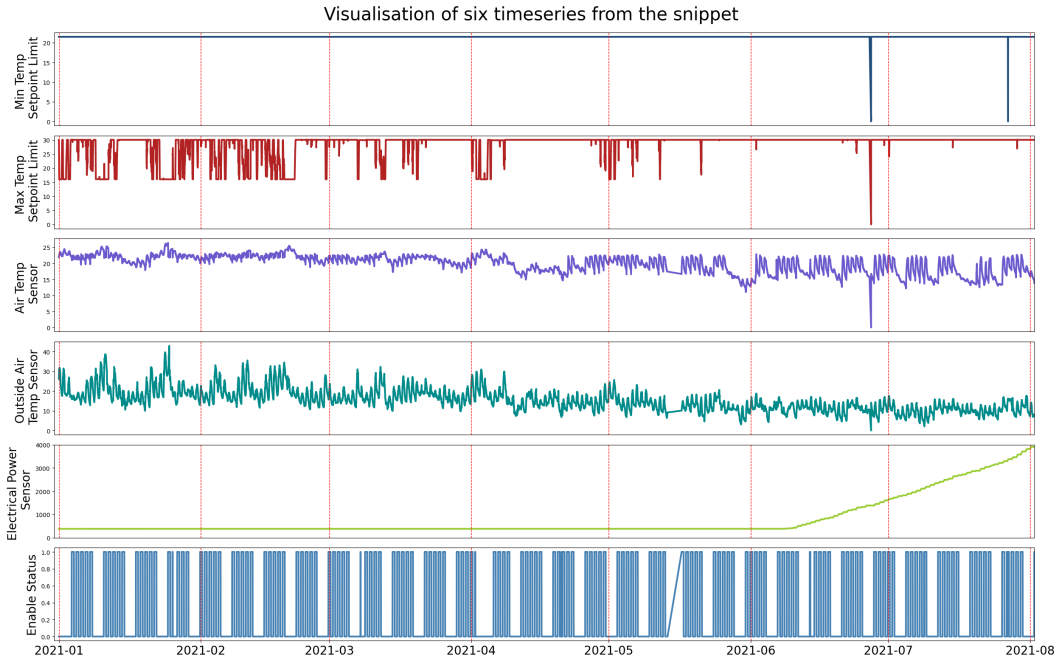


Figure 1: Visualisation of six timeseries with varying classes. The data is from the snippet of our BTS dataset available at `https://github.com/cruiseresearchgroup/DIEF_BTS`

# 3 Dataset

## 3.1 Collection Process

This dataset is comprised of data collected onto CSIRO's Data Clearing House (DCH `https://research.csiro.au/dch/`) digital platform [38]. Connecting to the Building Management Systems (BMS), timeseries data is collected from sensors, power, water and gas meters, and other devices within the buildings and uploaded using Message Queuing Telemetry Transport Secured (MQTTS). A semantic model of the building was created using DCH platform tooling. This created Brick schema [7] class definitions (version 1.2.1) for points within the model, and linked these points to the timeseries data ingested via MQTTS.

All instrumentation was conducted prior to the study, and as such no equipment installation or hardware setup was required by the authors. The work integrates with DCH platform which provides digital infrastructure to house building data, as well as to generate semantic models to describe the topology and instrumentation installed within the building. Based on a previously conducted systemic evaluation of existing ontologies suitable for our research context, we chose the Brick schema [66]. In terms of effort to map to the Brick schema, once sufficient details about the building are compiled, then typically expert engineers requires at least one to two days of per building to generate a full semantic building model.

Identifiers for both the point within the model, and the timeseries identifier were anonymised by generating Universally Unique Identifiers (UUID), and a three-year-period subset of the timeseries data was extracted from the DCH platform to produce this dataset. The data was not cleaned in effort to allow evaluation of various different cleaning algorithm, and to allow the evaluations of algorithms against data with realistic errors.

## 3.2 Description

**The Building TimeSeries (BTS) dataset** provides comprehensive, real-world data on building operations from three buildings in undisclosed Australian locations. It includes timeseries data (visualized in Figure 1) and building metadata standardised according to Brick schema [7]. Table 2


https://doi.org/10.52202/079017-4232

shows the statistics, comparing it to the LBNL59 dataset which is the only comparable dataset currently available. Part of this dataset have been presented in [45].

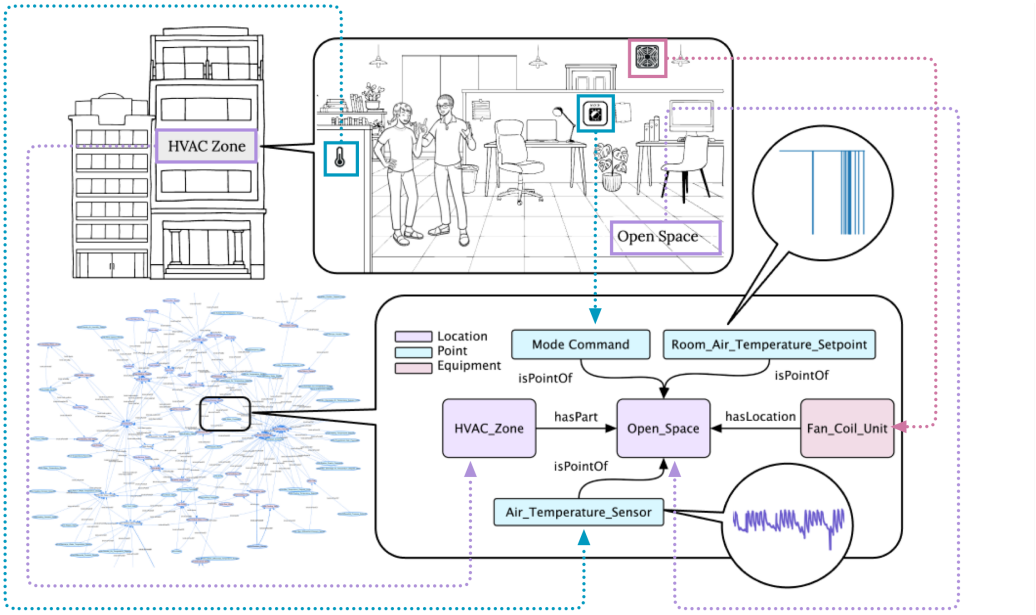


Figure 2: **Brick Schema Illustration and Visualization,** depicting machine-readable metadata for buildings as a knowledge graph. It reveals the logical and spatial links between distinct entities within a building, including the associated timeseries.

**Our dataset use the Brick schema**, a knowledge graph (KG) that details building components and their logical and spatial relationships. As illustrated in Figure 2, it specifies the equipment present in the buildings, the sensors attached to these equipment, their locations, and other related components within the same vicinity. Moreover, it also standardised the categorisations of the timeseries data into classes. The formal definition of the KG is as follows:

### 3.3 Formal definition of a building semantic model

A building contains many different entities, such as equipment in various locations, and these entities are interconnected. A structure that captures this information is called a "building semantic model" and can be interpreted as a KG. The mathematical formalisation of the "building semantic model" is a directed acyclic graph $\mathcal{G} = (V, P, E)$ where:

**Vertices (V)**: Each vertex $v \in V$ represents an entity within the building. This could be a physical location (e.g., a room or a zone served by a single HVAC subsystem), a piece of equipment (e.g. an air temperature sensor or a fan coil unit), or a reference to a time series in the form of a unique key. The actual time series data is typically stored in a separate database.

**Edges (E)**: Each edge $e = (u, p, v) \in E$ represents a predicate $p$ between two vertices $u$ and $v$.

**Predicate (P)**: Each edge $e$ is associated with a predicate $p \in P$ that specifies the type of relationship it represents (e.g., hasPart, has Location, or isPointOf).

#### 3.3.1 BTS and LBNL59

**BTS complements LBNL59** due to differences in time and location, as well as the size and complexity of the buildings. While LBNL59 covers a period ending in 2020 in the USA, our dataset spans from 2021 onwards in Australia, offering insights into longitudinal change and different seasonal patterns. Additionally, our dataset includes larger and more complex buildings compared to those in LBNL59.

[5]The reason for the discrepancy between the number of timeseries and `Point` is that multiple time series can be associated with the same Point in some instances.

Table 2: **Summary statistics** of the three buildings in our Building TimeSeries (BTS) dataset in comparison with LBNL59 [36, 49]. The table details the count and unique count (in parentheses) for the top-level Brick ontology [7] and the `Point` sub-classes. [5]

| | Count (Unique) | LBNL59 | | BTS_A | | BTS_B | | BTS_C | |
|---|---|---|---|---|---|---|---|---|---|
| Top Level | `Collection` | 0 | (0) | 4 | (2) | 2 | (2) | 8 | (1) |
| | `Equipment` | 59 | (3) | 547 | (24) | 159 | (25) | 963 | (41) |
| | `Location` | 73 | (3) | 481 | (9) | 68 | (17) | 381 | (26) |
| | `Point` | 230 | (11) | 8374 | (126) | 851 | (57) | 10440 | (159) |
| | Timeseries | 337 | | 8349 | | 851 | | 5347 | |
| Point Subclass | `Alarm` | 0 | (0) | 798 | (16) | 5 | (2) | 109 | (8) |
| | `Command` | 0 | (0) | 363 | (6) | 97 | (5) | 785 | (13) |
| | `Parameter` | 0 | (0) | 79 | (6) | 36 | (2) | 935 | (17) |
| | `Sensor` | 144 | (8) | 4396 | (56) | 266 | (25) | 4062 | (68) |
| | `Setpoint` | 86 | (3) | 772 | (26) | 232 | (16) | 1629 | (41) |
| | `Status` | 0 | (0) | 1628 | (17) | 110 | (6) | 2187 | (19) |
| | Location | Berkeley, USA | | Undisclosed locations in Australia | | | | | |
| | Start Date | 01-01-2018 | | 01-01-2021 | | 01-01-2021 | | 23-06-2021 | |
| | End Date | 31-12-2020 | | 31-12-2023 | | 31-12-2023 | | 18-01-2024 | |
| | Duration (Days) | 1094 | | 1094 | | 1094 | | 939 | |
| | Size Zipped (GB) | 0.26 | | 8.48 | | 1.31 | | 8.98 | |

**BTS dataset is larger and more diverse.** Each building in BTS includes significantly more timeseries—ranging from double to over twenty times more—resulting in a combined file size approximately 70 times larger when zipped.

The BTS dataset also exhibits greater diversity. Although LBNL59 contains 337 different timeseries, they are composed of only 11 different classes, all classified as either `Sensor` or `Setpoint`. In contrast, the BTS dataset has hundreds of unique `Point` classes including additional categories such as `Alarm`, `Command`, `Parameter`, and `Status`, offering a more comprehensive and varied dataset.

### 3.3.2 Addressing Literature Gaps with BTS Dataset

In Sections 2.2 and 2.3, the importance of scalability and interoperability was underscored, alongside the notable properties exhibited by our datasets, including domain shift, multimodality, imbalance, and long-tailedness. Here, we elaborate on how the BTS dataset effectively addresses these identified gaps in the literature.

**Brick is machine-readable and multimodal.** Consequently, this dataset fuels the research into building-agnostic, interoperable, and scalable software and ML models for building analytics. As a KG, Brick includes text components, facilitating novel research into interactions between KG, LLM and MVTS data.

**Our dataset is from real-world buildings.** This inclusion highlights real-world issues, as illustrated in Figure 1. For instance, the anomalously straight segments in `Air Temp Sensor`, `Outside Air Temp Sensor`, and `Enable Status` during the middle of May might indicate that there are missing values. Additionally, at the end of June, an anomalous data point is observed where the temperature sensors and setpoint limits drop to zero at the same time. It remains unclear if this was intentional, or by accident, or an error. This dataset serves as a test bed to evaluate how ML pipelines can address such issues during inference.

**Domain Shift.** The presence of domain shift complicates transfer learning efforts, as each building exhibits a unique distribution of classes. For instance, in the BTS_A, over half of the timeseries are `sensors`, whereas in BTS_B, this proportion drops to less than a third. Similarly, approximately a third of timeseries in BTS_B are `setpoints`, compared to less than a tenth in BTS_A.

Moreover, individual timeseries within each building demonstrate distinct distributions. As depicted in Figure 1, `Outside Air Temp Sensor` exhibit periodic behavior, leading to a more normal distribution, while `Electrical Power Sensor` display a non-periodic, monotonically increasing

 https://doi.org/10.52202/079017-4232

pattern, and `Enable Status` adheres to a Bernoulli distribution. Moreover, as shown in the figures in Appendix B, there is a significant disjoint of ontological classes between buildings; more than half of the classes only appear in one of the buildings only. Therefore, our dataset serves as an ideal dataset for investigating domain shifts.

**Long-Tailed Distributions.** The class distribution in BTS exhibits a long tail as shown in the figures in Appendix B. This means that certain class appear frequently, such as the 1004 instances of `Electrical Power Sensor` across all three buildings (Figure 4), while others are rare, with 10 classes appearing only once in the entire dataset, such as the `Air Differential Pressure Setpoint` location in `BTS_C` (Figure 7). Similarly, the values in some timeseries also follow a long-tailed distribution. For example, `Alarms` are expected to remain at zero most of the time.

## 4 Benchmark

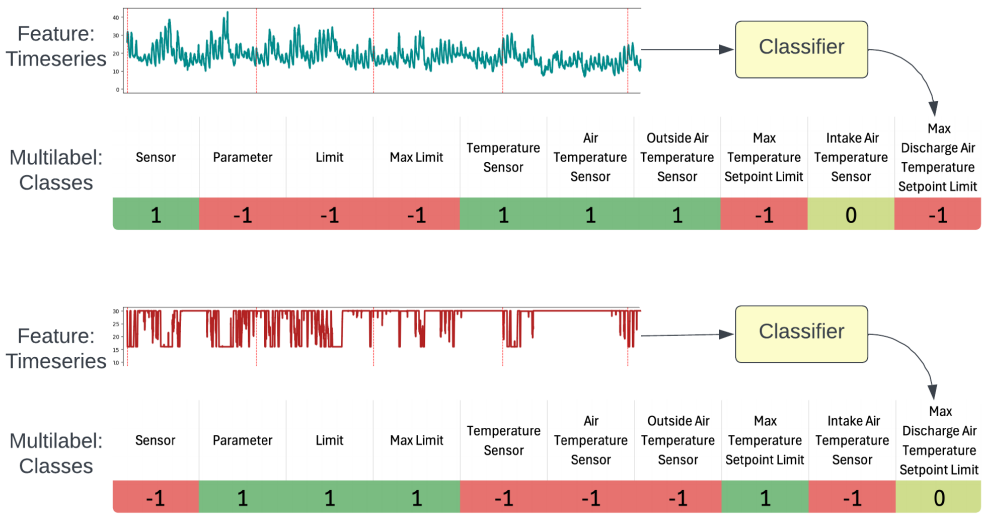


Figure 3: Visualisation of the multi-label timeseries classification task.

To demonstrate the utility of this dataset, we conducted a benchmark on the multi-label timeseries classification task. We picked this task because it highlights the challenges in implementing machine learning model that is interoperable between buildings. We also performed an additional benchmark on a zero-shot forecasting task. The details can be found in Appendix D.

Brick schema [7] was developed to aids in data interoperability across buildings. However, constructing the Brick schema for each building requires expensive and error prone manual expert labor to classifying timeseries data into the correct Brick classes. Past studies [8, 43, 67] have attempted to automate this process with ML relied on private data and did not release their code. This benchmark is the first to address the task using publicly available data. We formulated this task as a multi-label timeseries classification task, where a label will also return true for all super-classes and return as zero for all subclass. More details on this benchmark can be found in Appendix C

### 4.1 Problem Formulation

A datapoint $d = (t, v)$ is an ordered pair where $t \in \mathcal{R}$ is time and and $v \in \mathcal{R}$ is the value. A timeseries $T = \{d_i | 1 \leq i \leq n\}$ is a vector of datapoint of length $n \in \mathcal{Z}^+$. The length of timeseries can varies.

The class `Point` in Brick has $m$ sub-classes, including both direct and indirect sub-classes. In the original dataset, each timeseries is only labeled with a single class. However, we reformulated this as

a multi-label classification task, where a label will also return true for all super-classes and return as zero for all subclass. More formally, $l_j \in \{-1, 0, 1\}$ for $1 \leq j \leq m$ where $l_j = 1$ if timeseries $T$ belongs to the $j^{th}$ subclass of `Point` and also for all of its super-class, $l_j = 0$ for all of its sub-class, and $l_j = -1$ otherwise. For practical purposes, $m$ is not the number of sub-classes of `Point` in the definition, but only those found in our dataset.

The task for each timeseries is to predict if timeseries $T$ belongs in the $j^{th}$ label $l_j = f(T) \, \forall j$. This is visualised in Figure 3.

### 4.2 Results

Table 3: Benchmark results on the multi-label timeseries classification task. Deterministic methods do not have standard deviation.

| Method | Accuracy | | F1 | | mAP | |
|---|---|---|---|---|---|---|
| Zero | 0.8484 | ±N/A | 0.0000 | ±N/A | 0.0000 | ±N/A |
| Mode | 0.8592 | ±N/A | 0.1296 | ±N/A | 0.0990 | ±N/A |
| Random Proportional | 0.8147 | ±0.0001 | 0.1487 | ±0.0002 | 0.1520 | ±0.0001 |
| Random Uniform | 0.4999 | ±0.0002 | 0.1813 | ±0.0002 | 0.1520 | ±0.0001 |
| One | 0.1516 | ±N/A | 0.2234 | ±N/A | 0.1516 | ±N/A |
| LR | 0.2366 | ±N/A | 0.0882 | ±N/A | 0.0497 | ±N/A |
| XGBoost | 0.8593 | ±N/A | 0.2697 | ±N/A | 0.2627 | ±N/A |
| Transformer (default) | 0.7807 | ±0.0139 | 0.3360 | ±0.0116 | 0.3171 | ±0.0078 |
| Transformer (HP tuned) | 0.8052 | ±0.0074 | 0.3615 | ±0.0079 | 0.3489 | ±0.0057 |
| Informer | 0.7627 | ±0.0010 | 0.3162 | ±0.0019 | 0.2849 | ±0.0030 |
| DLinear | 0.7030 | ±0.0042 | 0.2499 | ±0.0020 | 0.2494 | ±0.0010 |
| PatchTST | 0.7534 | ±0.0017 | 0.2981 | ±0.0014 | 0.2721 | ±0.0013 |

Table 3 shows the results. Notice how naive methods achieved very high accuracy but very poor F1 and mean Average Precision (mAP) scores, while deep learning methods obtained slightly better F1 and mAP scores but much poor accuracy. We attribute this to the extreme imbalance in our dataset. All models performed only slightly better than the naive methods, indicating that this is an unsolved problem with significant potential for new discoveries.

Refer to Appendix C for more details about this experiment, including formal problem formulation, more results and other experimental details.

## 5 Limitations

Firstly, the dataset is sourced from only three non-residential buildings in Australia, limiting its geographical diversity. Consequently, models trained on this dataset may not generalize well to residential buildings, or buildings in other regions with different climates, regulations, and building practices. This limitation implies that models should primarily be used for research purposes rather than direct deployment.

Secondly, the anonymization process, essential for privacy, may have removed valuable context-specific information, such as building layouts, occupancy patterns, and operational schedules. This reduction in detail could limit the dataset's applicability for certain analyses. Moreover, despite thorough anonymization efforts, there is no absolute guarantee that personally identifiable information cannot be recovered, particularly when correlated with external datasets.

Finally, as this paper focuses on the dataset rather than benchmarking, the depth of the benchmarks is limited. For example, hyperparameter optimization was not performed.

## 6 Conclusion

In this paper, we introduced the Building TimeSeries (BTS) dataset, addressing the critical gaps in building analytics research by providing a comprehensive, publicly available dataset that spans three buildings over three years, encompassing over ten thousand timeseries data points and 240

 https://doi.org/10.52202/079017-4232

unique classes. This dataset is standardized using the Brick schema, ensuring interoperability and consistency across analyses. Additionally, our datasets inherently possess properties of interest to machine learning research, such as domain shift, multimodality, imbalance, and long-tailedness. Our benchmarks on multi-label timeseries classification and zero-shot forecasting tasks demonstrate the dataset's utility in addressing key challenges in building analytics. By making the BTS dataset and our benchmarking code publicly accessible, we aim to facilitate further research in optimizing building performance, ultimately contributing to efforts to mitigate climate change and enhance human well-being.

## Acknowledgments and Disclosure of Funding

This research is supported by the NSW Government through the CSIRO's NSW Digital Infrastructure Energy Flexibility (DIEF) project, funded under the Net Zero Plan Stage 1: 2020-2030.

This project is also funded by the Reliable Affordable Clean Energy for 2030 (RACE for 2030) Cooperative Research Centre.

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

## References

[1] U. Ali, M. H. Shamsi, M. Bohacek, K. Purcell, C. Hoare, E. Mangina, and J. O'Donnell. A data-driven approach for multi-scale gis-based building energy modeling for analysis, planning and support decision making. *Applied Energy*, 279:115834, 2020. 3

[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). Building energy models (bem) for commercial buildings based on cbecs data. `https://www.colorado.edu/lab/sbs/BEM`, 2021. 3, 19

[3] I. B. Arief-Ang, M. Hamilton, and F. D. Salim. A scalable room occupancy prediction with transferable time series decomposition of co2 sensor data. *ACM Transactions on Sensor Networks (TOSN)*, 14(3-4):1–28, 2018. 4

[4] I. B. Arief-Ang, F. D. Salim, and M. Hamilton. Da-hoc: semi-supervised domain adaptation for room occupancy prediction using co2 sensor data. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, pages 1–10, 2017. 4

[5] Ausgrid. Solar home electricity data. `https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data`, 2013. 2, 3, 19

[6] T. Babaei, H. Abdi, C. P. Lim, and S. Nahavandi. A study and a directory of energy consumption data sets of buildings. *Energy and Buildings*, 94:91–99, 2015. 3

[7] B. Balaji, A. Bhattacharya, G. Fierro, J. Gao, J. Gluck, D. Hong, A. Johansen, J. Koh, J. Ploennigs, Y. Agarwal, et al. Brick: Metadata schema for portable smart building applications. *Applied energy*, 226:1273–1292, 2018. 2, 5, 7, 8

[8] B. Balaji, C. Verma, B. Narayanaswamy, and Y. Agarwal. Zodiac: Organizing large deployment of sensors to create reusable applications for buildings. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, BuildSys '15, page 13–22, New York, NY, USA, 2015. Association for Computing Machinery. 8

[9] F. Biljecki. Exploration of open data in southeast asia to generate 3d building models. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, VI-4/W1-2020:37–44, 2020. 3, 19

[10] F. Biljecki and Y. S. Chow. Global building morphology indicators. *Computers, Environment and Urban Systems*, 95:101809, 2022. 3, 19

[11] D. Blum, J. Arroyo, S. Huang, J. Drgoňa, F. Jorissen, H. T. Walnum, Y. Chen, K. Benne, D. Vrabie, M. Wetter, et al. Building optimization testing framework (boptest) for simulation-based benchmarking of control strategies in buildings. *Journal of Building Performance Simulation*, 14(5):586–610, 2021. 4

[12] D. Blum, J. Candanedo, Z. Chen, G. Fierro, V. Gori, H. Johra, H. Madsen, A. Marszal-Pomianowska, Z. O'Neill, O. Pradhan, D. Rovas, F. Sacco, S. Stensson, C. A. Thilker, C. Vallianos, J. Wen, and S. White. *Data-Driven Smart Buildings: State-of-the-Art Review*. CSIRO, Australia, 2023. 1

[13] M. J. Booysen. Synthetic domestic hot water profile generator. *Stellenbosch University*, 1 2021. 3

[14] P. S. Dataport. Pecan street dataport. `https://dataport.pecanstreet.org/`, 2016. 3, 19

[15] S. Deldari, H. Xue, A. Saeed, J. He, D. V. Smith, and F. D. Salim. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint arXiv:2206.02353*, 2022. 4

[16] H. Deng, D. Fannon, and M. J. Eckelman. Predictive modeling for us commercial building energy use: A comparison of existing statistical and machine learning algorithms using cbecs microdata. *Energy and Buildings*, 163:34–43, 2018. 3, 19

[17] Department of City Planning. `https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page`. 3, 19

[18] B. Dong, Y. Liu, W. Mu, Z. Jiang, P. Pandey, T. Hong, B. Olesen, T. Lawrence, Z. O'Neil, C. Andrews, et al. A global building occupant behavior database. *Scientific data*, 9(1):369, 2022. 2, 3, 19

[19] S. Dooley, G. S. Khurana, C. Mohapatra, S. V. Naidu, and C. White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[20] J. Drgoňa, A. R. Tuor, V. Chandan, and D. L. Vrabie. Physics-constrained deep learning of multi-zone building thermal dynamics. *Energy and Buildings*, 243:110992, 2021. 3, 19

[21] ecobee. Donate your data. *ecobee*, 2019. 3, 19

[22] G. Fierro, M. Pritoni, M. AbdelBaky, D. Lengyel, J. Leyden, A. Prakash, P. Gupta, P. Raftery, T. Peffer, G. Thomson, et al. Mortar: an open testbed for portable building analytics. *ACM Transactions on Sensor Networks (TOSN)*, 16(1):1–31, 2019. 3, 19

[23] J. Gao, J. Ploennigs, and M. Berges. A data-driven meta-data inference framework for building automation systems. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, pages 23–32, 2015. 3, 19

[24] N. Gao, M. Marschall, J. Burry, S. Watkins, and F. D. Salim. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data*, 9(1):261, 2022. 3

[25] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), sep 2020. 3

[26] N. Gao, W. Shao, M. S. Rahaman, J. Zhai, K. David, and F. D. Salim. Transfer learning for thermal comfort prediction in multiple cities. *Building and Environment*, 195:107725, 2021. 4

[27] M. González-Torres, L. Pérez-Lombard, J. F. Coronel, I. R. Maestre, and D. Yan. A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Reports*, 8:626–637, 2022. 1

[28] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024. 2

 https://doi.org/10.52202/079017-4232

[29] H. B. Gunay, M. Ouf, G. Newsham, and W. O'Brien. Sensitivity analysis and optimization of building operations. *Energy and Buildings*, 199:164–175, 2019. 3, 19

[30] H. B. Gunay, W. Shen, G. Newsham, and A. Ashouri. Modelling and analysis of unsolicited temperature setpoint change requests in office buildings. *Building and Environment*, 133:203–212, 2018. 3, 19

[31] J. Haberl and J. Kreider. Instructions for" the great energy predictor shootout ii: Measuring retrofit energy savings", 1994. 3

[32] J. Haberl and J. Kreider. Predicting building energy usage: The great energy predictor shootout: Overview and discussion of results, 1994. 3

[33] F. Haldi and D. Robinson. Adaptive actions on shading devices in response to local visual stimuli. *Journal of Building Performance Simulation*, 3(2):135–153, 2010. 3, 19

[34] D. Hong, J. Ortiz, K. Whitehouse, and D. Culler. Towards automatic spatial verification of sensor placement in buildings. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8, 2013. 3, 19

[35] D. Hong, H. Wang, J. Ortiz, and K. Whitehouse. The building adapter: Towards quickly applying building analytics at scale. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, BuildSys '15, page 123–132, New York, NY, USA, 2015. Association for Computing Machinery. 3, 19

[36] T. Hong, N. Luo, D. Blum, and Z. Wang. A three-year building operational performance dataset for informing energy efficiency. `https://datadryad.org/stash/dataset/doi:10.7941/D1N33Q`, 2022. 1, 2, 3, 7, 19

[37] A. Howard, C. Balbach, C. Miller, J. Haberl, K. Gowri, and S. Dane. Ashrae - great energy predictor iii, 2019. 3

[38] D. Hugo, J. McCulloch, A. Hameed, W. Borghei, M. Grimeland, V. Felstead, and M. Goldsworthy. A smart building semantic platform to enable data re-use in energy analytics applications: the data clearing house, 2023. 5

[39] X. Jin, C. Fu, H. Kazmi, A. Balint, A. Canaydin, M. Quintana, F. Biljecki, F. Xiao, and C. Miller. The building data genome directory – an open, comprehensive data sharing platform for building performance research. *Journal of Physics: Conference Series*, 2600(3):032003, nov 2023. 3, 19

[40] X. Jin, C. Zhang, F. Xiao, A. Li, and C. Miller. A review and reflection on open datasets of city-level building energy use and their applications. *Energy and Buildings*, 285:112911, 2023. 3, 19

[41] M. Killian and M. Kozek. Ten questions concerning model predictive control for energy efficient buildings. *Building and Environment*, 105:403–412, 2016. 3

[42] J. Koh, B. Balaji, D. Sengupta, J. McAuley, R. Gupta, and Y. Agarwal. Scrabble: transferrable semi-automated semantic metadata normalization using intermediate representation. In *Proceedings of the 5th Conference on Systems for Built Environments*, pages 11–20, 2018. 3, 19

[43] J. Koh, D. Hong, R. Gupta, K. Whitehouse, H. Wang, and Y. Agarwal. Plaster: an integration, benchmark, and development framework for metadata normalization methods. In *Proceedings of the 5th Conference on Systems for Built Environments*, BuildSys '18, page 1–10, New York, NY, USA, 2018. Association for Computing Machinery. 3, 8, 19

[44] H. Li, H. Johra, F. de Andrade Pereira, T. Hong, J. Le Dréau, A. Maturo, M. Wei, Y. Liu, A. Saberi-Derakhtenjani, Z. Nagy, A. Marszal-Pomianowska, D. Finn, S. Miyata, K. Kaspar, K. Nweye, Z. O'Neill, F. Pallonetto, and B. Dong. Data-driven key performance indicators and datasets for building energy flexibility: A review and perspectives. *Applied Energy*, 343:121217, 2023. 3, 19

[45] X. Lin, A. Prabowo, I. Razzak, H. Xue, M. Amos, S. Behrens, S. White, and F. D. Salim. A gap in time: The challenge of processing heterogeneous iot point data in buildings. *arXiv preprint arXiv:2405.14267*, 2024. 2, 4, 6

[46] Y.-W. Lin, T. L. E. Tang, and C. J. Spanos. Hybrid approach for digital twins in the built environment. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, pages 450–457, 2021. 4

[47] Y. Liu, B. Dong, T. Hong, B. Olesen, T. Lawrence, and Z. O'Neill. Ashrae urp-1883: Development and analysis of the ashrae global occupant behavior database. *Science and Technology for the Built Environment*, 29(8):749–781, 2023. 2, 3, 19

[48] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023. 25, 26, 27

[49] N. Luo, Z. Wang, D. Blum, C. Weyandt, N. Bourassa, M. A. Piette, and T. Hong. A three-year dataset supporting research on building energy management and occupancy analytics. *Scientific data*, 9(1):156, 2022. 1, 2, 3, 7, 19

[50] T. Marzullo, S. Dey, N. Long, J. Leiva Vilaplana, and G. Henze. A high-fidelity building performance simulation test bed for the development and evaluation of advanced controls. *Journal of Building Performance Simulation*, 15(3):379–397, 2022. 4

[51] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J. Y. Park, Z. Nagy, P. Raftery, B. W. Hobson, Z. Shi, and F. Meggers. The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition. *Scientific Data*, 7:368, Oct. 2020. 2, 3, 19

[52] C. Miller and F. Meggers. The building data genome project: An open, public data set from non-residential building electrical meters. *Energy Procedia*, 122:439 – 444, 2017. {CISBAT} 2017 International ConferenceFuture Buildings & Districts – Energy Efficiency from Nano to Urban Scale. 3, 19

[53] Milojevic-Dupont, Nikola and Wagner, Felix, F. Nachtigall, J. Hu, G. B. Brüser, M. Zumwald, F. Biljecki, N. Heeren, L. H. Kaack, P.-P. Pichler, and F. Creutzig. Eubucco v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. *Scientific Data*, 10(1):147, 2023. 3, 19

[54] G. Z. Nagy. The CityLearn Challenge 2021. `https://doi.org/10.18738/T8/Q2EIQC`, 2021. 3, 19

[55] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. 25, 26, 27

[56] K. Nweye, M. Langtry, R. Choudhary, and G. Z. Nagy. The CityLearn Challenge 2023 Dataset. `https://doi.org/10.18738/T8/SXFWTI`, 2024. 3, 19

[57] K. Nweye, S. Siva, and G. Z. Nagy. The CityLearn Challenge 2022. `https://doi.org/10.18738/T8/0YLJ6Q`, 2023. 3, 19

[58] A. Parker, H. Horsey, M. Dahlhausen, M. Praprost, C. CaraDonna, A. LeBar, and L. Klun. Comstock reference documentation: Version 1. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2023. 3, 19

[59] L. Pereira and N. Nunes. Performance evaluation in non-intrusive load monitoring: datasets, metrics, and tools—a review. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 8(6):e1265, 2018. 3, 19

[60] A. Prabowo, K. Chen, H. Xue, S. Sethuvenkatraman, and F. D. Salim. Continually learning out-of-distribution spatiotemporal data for robust energy forecasting. In G. De Francisci Morales, C. Perlich, N. Ruchansky, N. Kourtellis, E. Baralis, and F. Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 3–19, Cham, 2023. Springer Nature Switzerland. 3, 19

[61] A. Prabowo, K. Chen, H. Xue, S. Sethuvenkatraman, and F. D. Salim. Navigating out-of-distribution electricity load forecasting during covid-19: A continual learning approach leveraging human mobility. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2023. 3, 19

[62] A. Prabowo, W. Shao, H. Xue, P. Koniusz, and F. D. Salim. Because every sensor is unique, so is every pair: Handling dynamicity in traffic forecasting. In *9th ACM/IEEE Conference on Internet of Things Design and Implementation (IoTDI)*, IoTDI '23, page 93–104, New York, NY, USA, 2023. Association for Computing Machinery. 4

[63] A. Prabowo, H. Xue, W. Shao, P. Koniusz, and F. D. Salim. Traffic forecasting on new roads unseen in the training data using spatial contrastive pre-training. *Data Mining and Knowledge Discovery*, 2023. 4

[64] T. Péan, R. Costa-Castelló, E. Fuentes, and J. Salom. Experimental testing of variable speed heat pump control strategies for enhancing energy flexibility in buildings. *IEEE Access*, 7:37071–37087, 2019. 3, 19

[65] T. Péan and J. Salom. Experimental HIL datasets of a heat pump controlled by MPC or rule-based controllers for energy flexibility. `https://doi.org/10.5281/zenodo.7006826`, Aug. 2022. 3, 19

[66] Z. Qiang, S. Hands, K. Taylor, S. Sethuvenkatraman, D. Hugo, P. Ghiasnezhad Omran, M. Perera, and A. Haller. A systematic comparison and evaluation of building ontologies for deploying data-driven analytics in smart buildings. *Energy and Buildings*, 292:113054, 2023. 5

[67] M. Rana, A. Rahman, M. Almashor, J. McCulloch, and S. Sethuvenkatraman. Automatic classification of sensors in buildings: Learning from time series data. In T. Liu, G. Webb, L. Yue, and D. Wang, editors, *AI 2023: Advances in Artificial Intelligence*, pages 367–378, Singapore, 2024. Springer Nature Singapore. 2, 3, 8, 19

[68] Regione Lombardia, Azienda Regionale per l'Innovazione e gli Acquisti (ARIA). Cened+2. `https://www.dati.lombardia.it/Energia/Database-CENED-2-Certificazione-ENergetica-degli-E/bbky-sde/about_data`, 2024. 3, 19

[69] H. B. Rijal, P. Tuohy, F. Nicol, M. A. Humphreys, A. Samuel, and J. Clarke. Development of an adaptive window-opening algorithm to predict the thermal comfort, energy use and overheating in buildings. *Journal of building performance simulation*, 1(1):17–30, 2008. 3, 19

[70] M. Ritchie, J. Engelbrecht, and M. Booysen. A probabilistic hot water usage model and simulator for use in residential energy management. *Energy and Buildings*, 235:110727, 2021. 3

[71] J. Rubio-Herrero, V. Chandan, C. Siegel, A. Vishnu, and D. Vrabie. A learning framework for control-oriented modeling of buildings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 473–478. IEEE, 2017. 3, 19

[72] F. D. Salim, B. Dong, M. Ouf, Q. Wang, I. Pigliautile, X. Kang, T. Hong, W. Wu, Y. Liu, S. K. Rumi, et al. Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. *Building and Environment*, 183:106964, 2020. 1, 2, 3, 19

[73] W. Shao, S. Zhao, Z. Zhang, S. Wang, M. S. Rahaman, A. Song, and F. D. Salim. Fadacs: A few-shot adversarial domain adaptation architecture for context-aware parking availability sensing. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2021. 4

[74] X. Shi, B. Si, J. Zhao, Z. Tian, C. Wang, X. Jin, and X. Zhou. Magnitude, causes, and solutions of the performance gap of buildings: A review. *Sustainability*, 11(3):937, 2019. 3

[75] Sustainable Energy Authority of Irleand. Ber tool. `https://ndber.seai.ie/BERResearchTool/ber/search.aspx`. 3, 19

[76] H. Syse and H. Nikpey. Building performance simulation of mybox energy lab in norway: Investigating the human dimension in energy use analysis. EasyChair Preprint no. 13521, EasyChair, 2024. 3, 19

[77] J. Taneja, A. Krioukov, S. Dawson-Haggerty, and D. Culler. Enabling advanced environmental conditioning with a building application stack. In *2013 International Green Computing Conference Proceedings*, pages 1–10, 2013. 3, 19

[78] A. Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86. 2, 3, 19

[79] UK Power Networks. Smartmeter energy consumption data in london households. `https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households`, 2015. 2, 3, 19

[80] O. o. S. U.S. Department of Energy and T. I. (OSTI). Energyplus™, version 00, 9 2017. 4

[81] J. Vazquez Canteli and Z. Nagy. The CityLearn Challenge 2020. `https://doi.org/10.18738/T8/ZQKK6E`, 2020. 3, 19

[82] P. Wei and X. Jiang. Data-driven energy and population estimation for real-time city-wide energy footprinting. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '19, page 267–276, New York, NY, USA, 2019. Association for Computing Machinery. 3

[83] M. Wetter, K. Benne, H. Tummescheit, and C. Winther. Spawn: coupling modelica buildings library and energyplus to enable new energy system and control applications. *Journal of Building Performance Simulation*, 17(2):274–292, 2024. 4

[84] Wilson, Eric, Parker, Andrew, Fontanini, Anthony, Present, Elaina, Reyna, Janet, Adhikari, Rajendra, Bianchi, Carlo, CaraDonna, Christopher, Dahlhausen, Matthew, Kim, Janghyun, LeBar, Amy, Liu, Lixi, Praprost, Marlena, White, Philip, Zhang, Liang, DeWitt, Peter, Merket, Noel, Speake, Andrew, Hong, Tianzhen, Li, Han, M. Frick, Natalie, Wang, Zhe, Blair, Aileen, Horsey, Henry, Roberts, David, Trenbath, Kim, Adekanye, Oluwatobi, Bonnema, Eric, E. Kontar, Rawad, Gonzalez, Jonathan, Horowitz, Scott, Jones, Dalton, Muehleisen, Ralph, Platthotam, Siby, Reynolds, Matthew, Robertson, Joseph, Sayers, Kevin, and Q. Li. End-use load profiles for the u.s. building stock. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 10 2021. 3, 19

[85] E. J. Wilson, A. Parker, A. Fontanini, E. Present, J. L. Reyna, R. Adhikari, C. Bianchi, C. CaraDonna, M. Dahlhausen, J. Kim, et al. End-use load profiles for the us building stock: Methodology and results of model calibration, validation, and uncertainty quantification. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2022. 3

[86] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*. PMLR, 2024. 4

[87] A. N. Wu and F. Biljecki. Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. *Landscape and Urban Planning*, 214:104167, 2021. 3, 19

[88] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. 22, 25

[89] H. Xue and F. D. Salim. Utilizing language models for energy load forecasting. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '23, page 224–227, New York, NY, USA, 2023. Association for Computing Machinery. 4

[90] Y. Ye, K. Hinkelman, J. Zhang, W. Zuo, and G. Wang. A methodology to create prototypical building energy models for existing buildings: A case study on us religious worship buildings. *Energy and Buildings*, 194:351–365, 2019. 3, 19

[91] Y. Ye, G. Wang, and W. Zuo. Creation of a prototype building model of college and university building. In *Proceedings of the 4th International Conference on Building Energy and Environment (COBEE2018), Melbourne, Australia*, 2018. 3, 19

[92] Y. Ye, G. Wang, W. Zuo, P. Yang, and K. Joshi. Development of a baseline building model of auto service and repair shop. In *Proceedings of 2018 ASHRAE Building Performance Analysis Conference and SimBuild (BPACS2018), Chicago, IL, USA*, 2018. 3, 19

[93] A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023. 25, 26, 27

[94] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024. 4

[95] H.-x. Zhao and F. Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012. 3, 4

[96] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021. 25, 26, 27

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 6.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included the URL in the abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix C and D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix C and D
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C and D. Not the total amount of compute, but it was not significant.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] No existing assets were used.
   (b) Did you mention the license of the assets? [Yes] The data is released under `CC BY 4.0` while the code is released under `MIT License`. These are mentioned in their respective repository.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included the URL in the abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The dataset does not contain personally identifiable information.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The dataset does not contain personally identifiable information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]


https://doi.org/10.52202/079017-4232

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendices

## A List of Related Works

For convenience, we summarised the works listed in Section 2.1 in Table 4.

Table 4: List of related work.

| | **Datasets** |
|---|---|
| Private | HVAC [69, 33, 77, 71, 30, 29, 20], energy use [60, 61], timeseries ontology classification [34, 35, 23, 42, 43, 67], and simulation [76]. |
| Paid | Pecan Street [14]. |
| Upon discretion of the data provider | ecobee [21]. Mortar [22] is intended to be freely available, yet it has limited access due to cloud deployment issues at the time of writing). |
| Static | EUBUCCO [53], PLUTO [17], GBMI [10], Roofpedia [87], HBD3D [9], |
| Corase temporal granularity (more than daily) | CBECS [16], BERTOOL [75], CENED+2 [68], |
| Simulation-based | BEM4CBECS [2, 91, 92, 90], ResStock [84], ComStock [58], CityLearn Challenge Series [81, 54, 57, 56], and hardware-in-the-loop laboratory [65, 64]. |
| Limited scope | SLRHOME [5], LCLD [79], and UCI [78] |
| NILM | Non-intrusive load monitoring (NILM) is task and many dataset have been made for this task check this recent survey [59] that list publicly available dataset. However, since the datasets are only made for this specific task in mind, the scope is limited to only electricity submetering. Other datasets with focus on submetering: BDG [52] and BDG2 [51]. |
| Occupant behaviour | From AshraeOB [18, 47] website: "The ASHRAE Global Occupant Behavior Database aims to advance the knowledge and understanding of realistic occupancy patterns and human-building interactions with building systems. This database includes 34 field-measured occupant behavior datasets for both commercial and residential buildings, contributed by researchers from 15 countries and 39 institutions covering 10 different climate zones. It includes occupancy patterns, occupant behaviors, indoor and outdoor environment measurements." |
| **Comprehensive** | Lawrence Berkeley National Laboratory building 59 (LBNL59) [36, 49] and **BTS (ours)** `https://github.com/cruiseresearchgroup/DIEF_BTS`. |
| Other lists | A review paper on NILM [59], a review paper on buildings at urban scale [72], a review paper on energy flexibility datasets [44], a review paper on building and energy dataset [40], and the Building Data Genome Directory [39]. |

## B Visualisation of Domain Shift and Long-tail Distribution in Our Datasets.

We visualise the domain shift by comparing the different distributions of classes between buildings. We visualise the that the distributions of classes have long-tails by plotting the histogram. These are shown in Figure 4, 5, 6, and 7. The relevant discussions can be found in Section 2.3 and 3.3.2.

 https://doi.org/10.52202/079017-4232

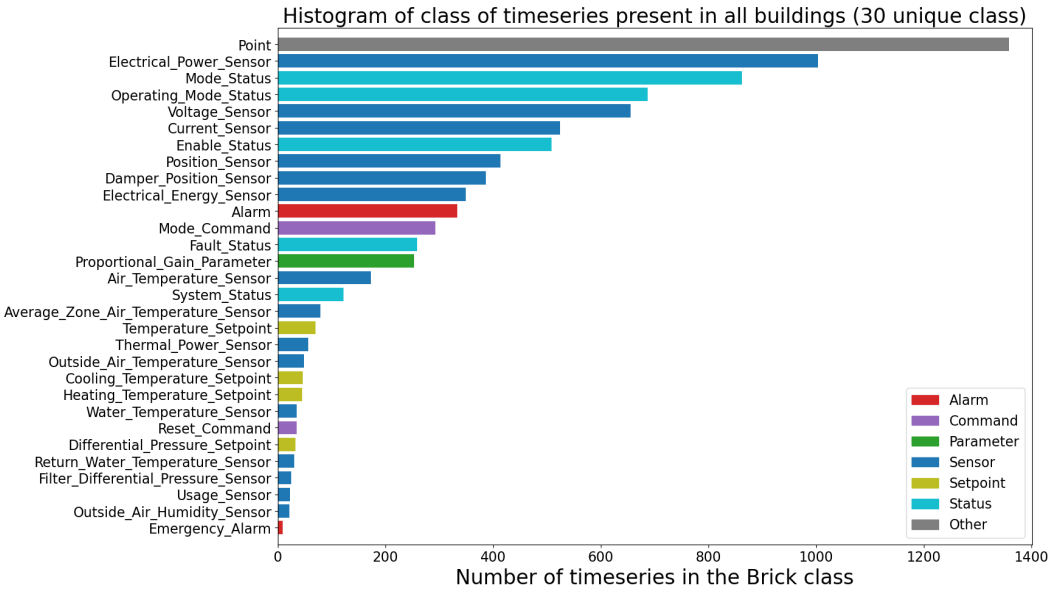


Figure 4: Histogram of class of timeseries by buildings.

# C Multi-label Timeseries Classification: More details

## C.1 Data Pre-processing

Each timeseries are cut into shorter chunk of either 2/4/8 weeks. The reason is to enable analysis of accuracy against various length of the timeseries. Those with too few datapoint, less than 1 per day, are removed. Due to great ranges of values, they are scaled using symmetric log first, and then standard scaling. The symmetric log function is defined as follows:

$$v' = \begin{cases} 9 + \log_{10}(v) & \text{if } v > 10 \\ -9 - \log_{10}(-v) & \text{if } v < -10 \\ v & \text{otherwise} \end{cases}$$

## C.2 Development and Test Partition

The partition is done by time and buildings. The reason for this partition strategy is to evaluate the performance in the future, and in different buildings. The development partition consist of the first four months of `BTC_A` and the first year of `BTC_B`. The development partition is randomly split into training and validation with a 80% and 20% ratio respectively. The remaining data are set to the testing partition.

## C.3 Feature Extraction

Depending on whether the models are made generic classification (LR, RF, and XGBoost) or deep learning models specialised for timeseries, a different feature extraction method were used. For generic models, we extract the following global features: mean, standard deviation, skew, kurtosis, root mean square, minimum, maximum, the three quartiles, and average duration between data points. For timeseries algorithm, we aggregate the timeseries into four hour slots and extract the maximum, mean, standard deviation, and number of datapoints within each slot.

## C.4 Model Training

We used binary cross-entropy (BCE) loss, treating every single label as binary, and applied additional extra weight to the positive samples proportionally. The maximum number of epochs was set to 100,

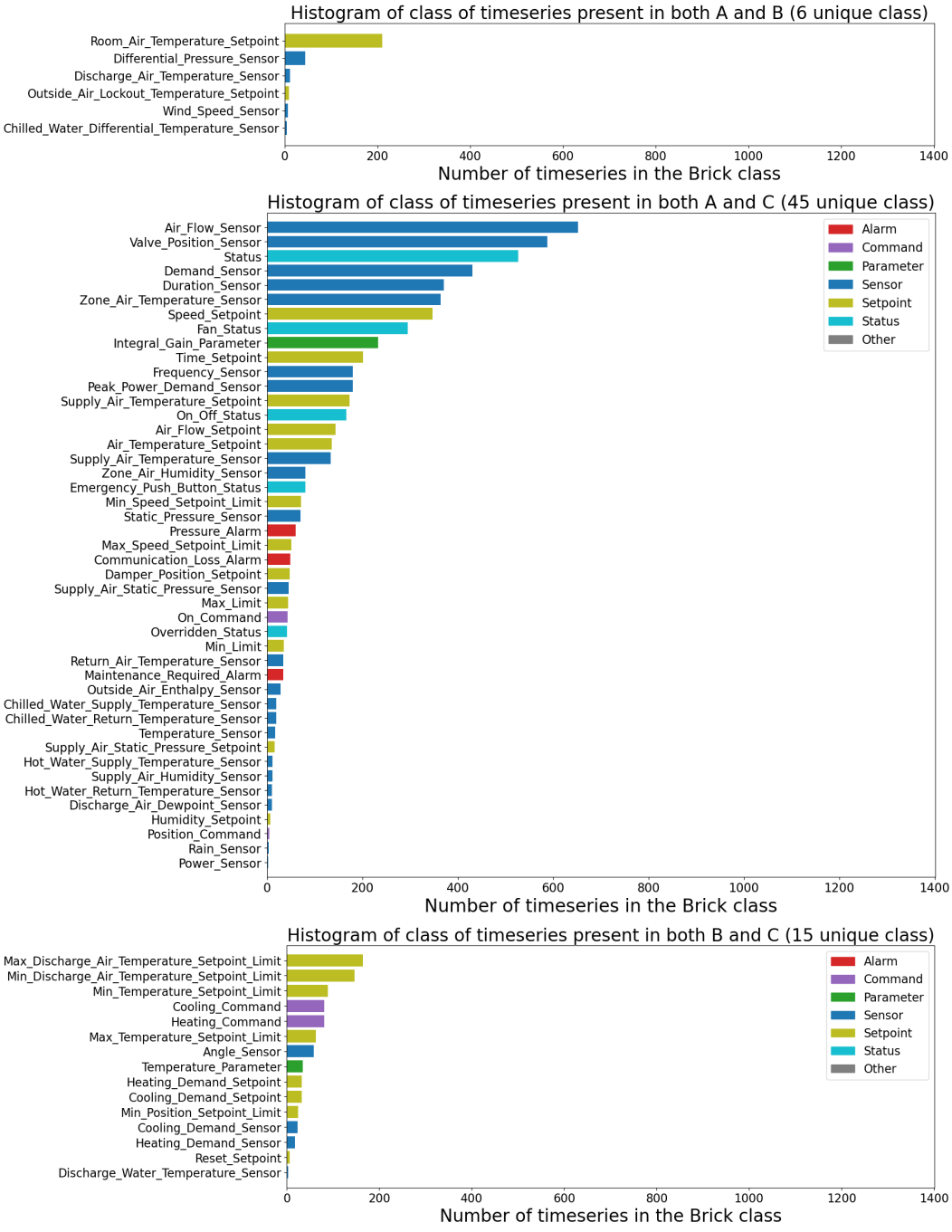


Figure 5: Histogram of class of timeseries by buildings, continued.


https://doi.org/10.52202/079017-4232

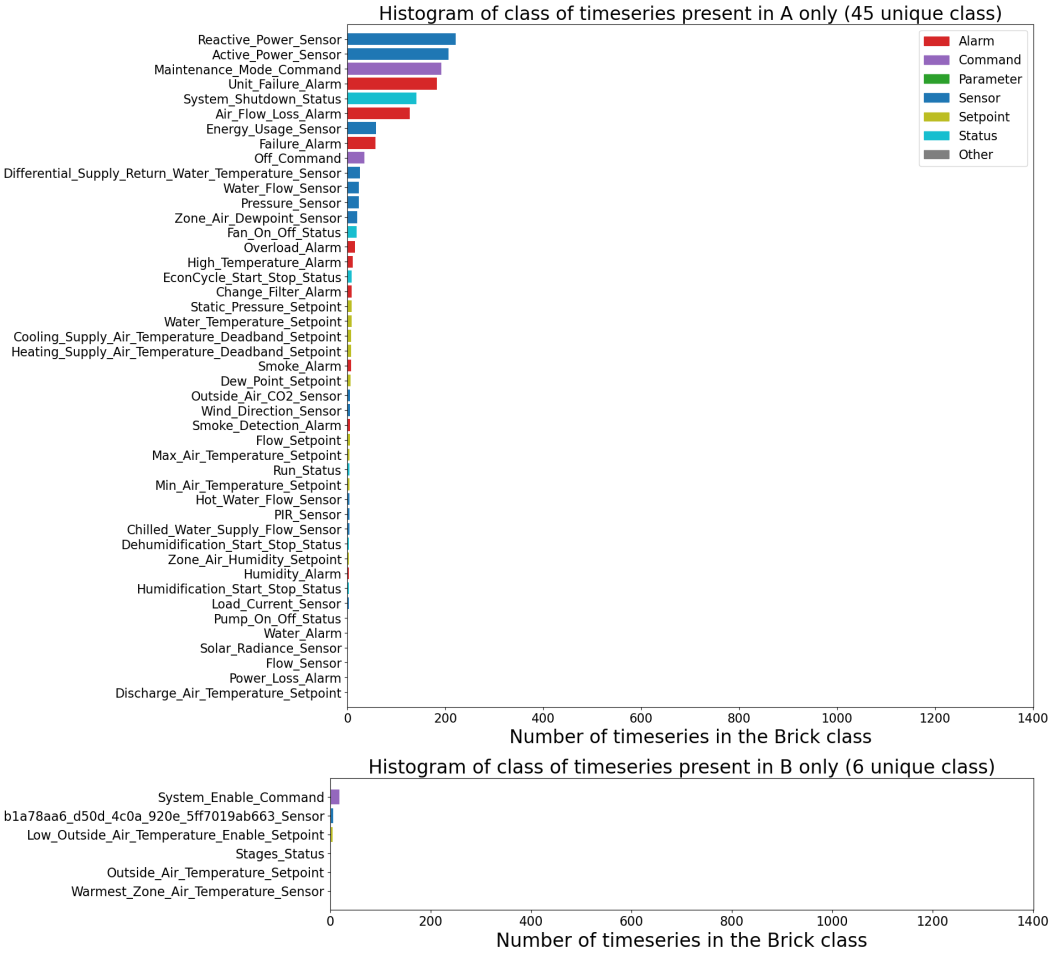


Figure 6: Histogram of class of timeseries by buildings, continued.

with a patience of 30 epochs for early stopping. The learning rate was set to 0.01, and we used the ReduceLROnPlateau strategy with a patience of 10 epochs. The optimizer was Rectified Adam (RAdam). For deep learning methods, we adapted the TSLib code [88] from their official GitHub repository `https://github.com/thuml/Time-Series-Library`. The batch size for each method was adjusted to fit memory. Our implementations, including our hardware setup, are available on the GitHub repository for this project `https://github.com/cruiseresearchgroup/DIEF_BTS`.

### C.5 Baselines

We use four naive baselines that does not take the features into account:

- **Zero.** The model output negative prediction on all labels.
- **Random Uniform.** The model based the prediction on a coin flip (50/50)
- **Random Proportional.** The model based the prediction randomly, but according to the proportion each label appears on the training data.
- **Mode.** The most common class was `Sensor`. So the model predict`Sensor` all the time.

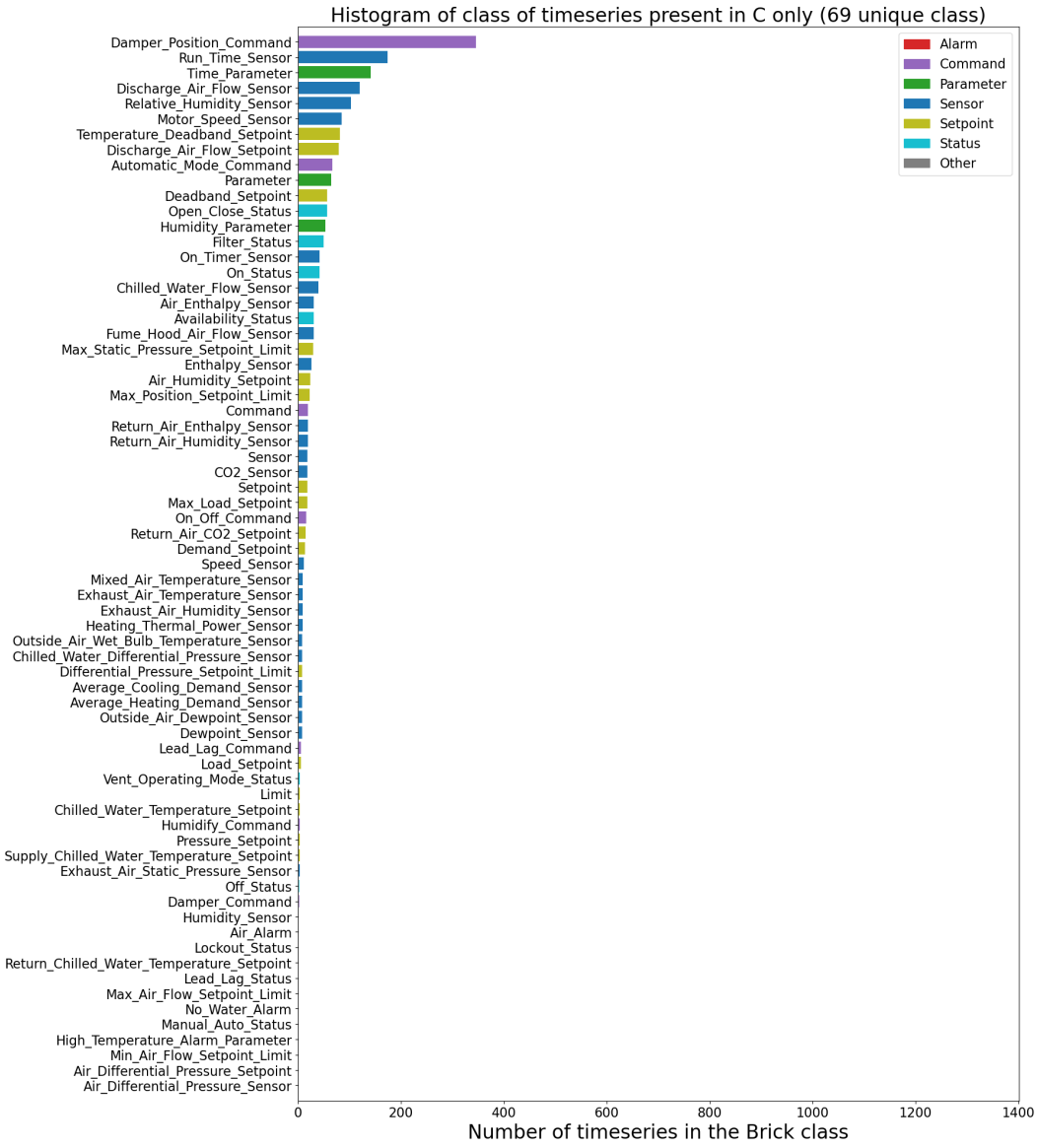


Figure 7: Histogram of class of timeseries by buildings, continued.


https://doi.org/10.52202/079017-4232

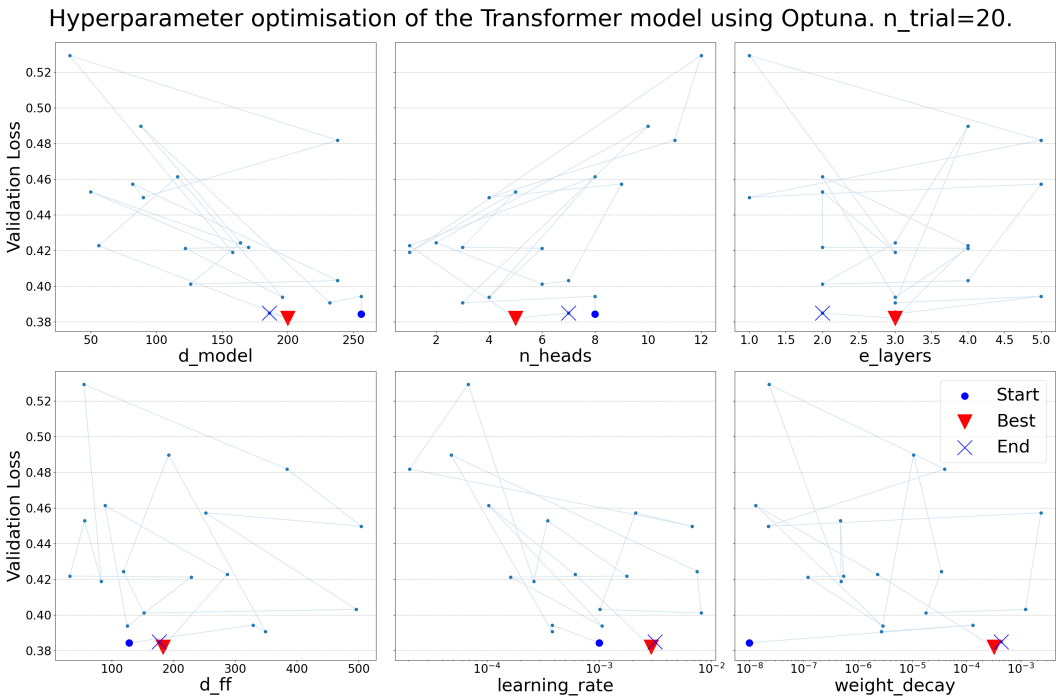


Figure 8: Visualisation of the hyperparameter tuning of the Transformer model.

### C.6 Hyperparameter Tuning

## D Zero-shot Forecasting Across Buildings

The advent of building digitalization presents significant opportunities for leveraging deep learning methods in building management systems for accurate forecasting. In practical applications, it is crucial for well-trained models to be applicable across diverse building scenarios without retraining costs. However, specific building constraints, operational variances, functionality differences, and data heterogeneity pose significant challenges in real-world settings. As shown in Table 2, models must adapt to dynamic ontology changes when applied to different buildings. Previous studies often rely on identical features and well-processed data, not reflecting the complexity of real-world scenarios. LBNL59, involving only one building, is insufficient for transfer learning studies. This study establishes a baseline for zero-shot forecasting using the BTS multivariate time series.

### D.1 Problem Formulation

Suppose we have dataset $D \in \mathcal{R}^{N \times K}$ with $N$ IoT points and $K$ timesteps. Each data point is denoted as $d_{N,k} = D_{N,k:k+S}$, where $S$ is the sequence length of the historical data. Detnote the forecasting model as $h(\cdot)$. The multi-step forecasting problem is formalized as follows: $h(d_{N,k}) = d_{N,k+S+H}$, where $H$ is the forecasting horizon. In zero-shot forecasting, the model is trained and tested across different datasets. In this study, $S = 12, H = 12$.

### D.2 Data Pre-processing

This study utilizes a 1-month training dataset spanning from 00:00:00 on 01/07/2022 to 00:00:00 on 01/08/2022, with irregular data resampled to a 10-minute granularity and then standardization. The historical window and forecast horizon are set to 12 time steps, equivalent to 2 hours. A model trained on one dataset is evaluated across all buildings for the same period. For each dataset, a subset of IoT points is selected for training based on the criterion $N_{\text{unique}}/N_{\text{sample}} > \eta$ where $\eta = 0.1$. This feature selection results in 133, 710, and 2025 IoT points for the three respective datasets.

### D.3 Baselines

DLinear[93], PatchTST[55], Informer[96] and iTransformer[48] as backbone models are employed for this benchmark study.

### D.4 Model Training

While employing DLinear for training, we treat this task as a multivariate forecasting task. Models are fed by all the IoT points data and expect to forecast the corresponding values of these IoT points. Considering that certain Transformer-based backbone models that involve a conventional embedding layer, such as iTransformer, Informer, and PatchTST, do not support changes in input channels between training and testing sets, we handle the task as an univariate forecasting problem, treating each IoT point equivalently. Similar to the multi-label classification task, the code is modified based on TSLib[88] Github repository. The training process employs the Adam optimizer with a learning rate of 0.01 and Mean Square Error (MSE) loss, and a learning rate scheduler is applied. Training is capped at 20 epochs with an early stopping patience of 3 epochs. All experiments are conducted on the NCI Gadi server utilizing 4 V100 GPUs.

### D.5 Metrics

Baseline performance is evaluated using Mean Absolute Error (MAE) and Symmetric Mean Absolute Percentage Error (SMPAE) averaged by IoT points.Following the above-mentioned notation, the mathematical definitions are as follows:

$$
\begin{aligned}
MAE &= \frac{1}{N}\sum_{n=1}^{N}|\hat{y_n} - y_n| \\
SMAPE &= \frac{100\%}{N}\sum_{n=1}^{N}\frac{|\hat{y_n} - y_n|}{|\hat{y_n}| + |y_n|} \\
R^2 &= 1 - \sum_{n=1}^{N}(y_n - \hat{y}_n)^2 / \sum_{n=1}^{N}(y_n - \bar{y})^2
\end{aligned}
$$

where $\hat{y}_n, y_n, \bar{y}_n$ are the multi-step prediction, ground truth, and mean for the evaluated model.

### D.6 Main Results

Table 5 presents the Symmetric Mean Absolute Percentage Error (SMAPE) and $R^2$ scores for four baseline models in this task, with diagonal values omitted. PatchTST and DLinear consistently outperform the other models, balancing higher $R^2$ scores with lower SMAPE values. However, the overall performance highlights the complexity and challenges inherent in zero-shot forecasting, indicating significant scope for further research and improvement.

### D.7 Detailed Results with Standard Deviations

Table6-8 shows the mean and standard deviation values about MAE, SMAPE, and $R^2$ for the multi-step zero-shot forecasting.

 https://doi.org/10.52202/079017-4232

Table 5: Benchmark results on the zero-shot forecasting task. The columns refer to the training set, whereas the row represents the testing set.

| | | BTS-A | | BTS-B | | BTS-C | |
|---|---|---|---|---|---|---|---|
| | | MAE | SMAPE | MAE | SMAPE | MAE | SMAPE |
| Previous Day Persistence | | 0.5377 | 48.1539 | 0.4976 | 43.2985 | 0.5458 | 45.7014 |
| Previous Week Persistence | | 0.6190 | 57.2713 | 0.5918 | 51.3867 | 0.6499 | 58.1922 |
| BTS-A | DLinear | N/A | | 0.4324 | 35.9846 | 0.4262 | 36.2734 |
| | PatchTST | N/A | | 0.3748 | 29.2570 | 0.3712 | 29.5552 |
| | Informer | N/A | | 0.5968 | 49.2217 | 0.5920 | 51.9745 |
| | iTransformer | N/A | | 0.4026 | 31.1924 | 0.3842 | 30.1102 |
| BTS-B | DLinear | 0.4940 | 41.2264 | N/A | | 0.4206 | 35.3121 |
| | PatchTST | 0.4575 | 36.7689 | N/A | | 0.3711 | 29.2135 |
| | Informer | 0.5233 | 45.9279 | N/A | | 0.4592 | 39.7068 |
| | iTransformer | 0.4783 | 37.5907 | N/A | | 0.3901 | 29.9940 |
| BTS-C | DLinear | 0.4858 | 40.7421 | 0.4158 | 34.1473 | N/A | |
| | PatchTST | 0.4542 | 36.9451 | 0.3723 | 28.9325 | N/A | |
| | Informer | 0.5213 | 46.6112 | 0.4602 | 39.7162 | N/A | |
| | iTransformer | 0.4859 | 39.5158 | 0.4262 | 32.6550 | N/A | |

Table 6: Mean Absolute Error (MAE) on the zero-shot forecasting task. The columns refer to the training set, whereas the row represents the testing set.

| | **Method** | **BTS_A** | **BTS_B** | **BTS_C** |
|---|---|---|---|---|
| **BTS_A** | DLinear[93] | N/A | 0.43243±0.16060 | 0.42617±0.19525 |
| | PatchTST [55] | N/A | 0.37480±0.06301 | 0.37480±0.06301 |
| | Informer [96] | N/A | 0.59679±0.04698 | 0.59196±0.05424 |
| | iTransformer [48] | N/A | 0.40257±0.06487 | 0.38416±0.07446 |
| **BTS_B** | DLinear[93] | 0.49398±0.21579 | N/A | 0.42059±0.20122 |
| | PatchTST [55] | 0.45745±0.08428 | N/A | 0.37106±0.07449 |
| | Informer [96] | 0.52329±0.06606 | N/A | 0.45922±0.05966 |
| | iTransformer [48] | 0.47830±0.08542 | N/A | 0.39099±0.07722 |
| **BTS_C** | DLinear[93] | 0.48582±0.22002 | 0.41582±0.17401 | N/A |
| | PatchTST [55] | 0.45413±0.08338 | 0.37227±0.06339 | N/A |
| | Informer [96] | 0.52133±0.06237 | 0.46022±0.05043 | N/A |
| | iTransformer [48] | 0.48588±0.08002 | 0.42620±0.06586 | N/A |

Table 7: Symmetric Mean Absolute Percentage Error (SMAPE) on the zero-shot forecasting task. The columns refer to the training set, whereas the row represents the testing set.

| | **Method** | **BTS_A** | **BTS_B** | **BTS_C** |
|---|---|---|---|---|
| **BTS_A** | DLinear[93] | N/A | 35.98461±15.47196 | 36.27335±18.34376 |
| | PatchTST [55] | N/A | 29.25704±5.03140 | 29.55517±6.07105 |
| | Informer [96] | N/A | 49.22169±2.54525 | 51.97452±4.25621 |
| | iTransformer [48] | N/A | 31.19242±5.23906 | 30.11023±5.97160 |
| **BTS_B** | DLinear[93] | 41.22638±18.84817 | N/A | 35.31209±18.23204 |
| | PatchTST [55] | 36.76894±6.63363 | N/A | 29.21348±5.96805 |
| | Informer [96] | 45.92792±6.15185 | N/A | 39.70681±5.37708 |
| | iTransformer [48] | 37.59074±6.54195 | N/A | 29.99402±6.02286 |
| **BTS_C** | DLinear[93] | 40.74205±19.53859 | 34.14733±16.12281 | N/A |
| | PatchTST [55] | 36.94508±6.74060 | 28.93252±5.03300 | N/A |
| | Informer [96] | 46.61115±6.07310 | 39.71622±4.55301 | N/A |
| | iTransformer [48] | 39.51578±6.64577 | 32.65497±5.24526 | N/A |

Table 8: $R^2$ score on the zero-shot forecasting task. The columns refer to the training set, whereas the row represents the testing set.

| | Method | BTS_A | BTS_B | BTS_C |
|---|---|---|---|---|
| **BTS_A** | DLinear[93] | N/A | 0.54196±0.12989 | 0.53206±0.09756 |
| | PatchTST [55] | N/A | 0.51219±0.16793 | 0.51258±0.05317 |
| | Informer [96] | N/A | 0.32122±0.18004 | 0.32153±0.05191 |
| | iTransformer [48] | N/A | 0.46723±0.17016 | 0.48543±0.05315 |
| **BTS_B** | DLinear[93] | 0.43686±0.09253 | N/A | 0.52964±0.09715 |
| | PatchTST [55] | 0.40926±0.03239 | N/A | 0.50624±0.05375 |
| | Informer [96] | 0.39893±0.02753 | N/A | 0.47109±0.04673 |
| | iTransformer [48] | 0.36844±0.03443 | N/A | 0.46792±0.05684 |
| **BTS_C** | DLinear[93] | 0.44519±0.09250 | 0.54543±0.12879 | N/A |
| | PatchTST [55] | 0.41773±0.03099 | 0.51411±0.17089 | N/A |
| | Informer [96] | 0.41886±0.02556 | 0.48993±0.13881 | N/A |
| | iTransformer [48] | 0.37250±0.03034 | 0.42437±0.17611 | N/A |

Models trained on BTS_A exhibit poorer cross-building forecasting results. This can be attributed to the greater complexity of BTS_A compared to BTS_B and BTS_C. BTS_A includes more heterogeneous series and entity types (BTS_A has 42 entities, where BTS_B and BTS_C have 16 and 31 entities respectively in the task training data), which introduces additional noise that impacts accuracy.

The evaluation metrics, MAE and SMAPE, indicate that PatchTST outperforms other baselines, while $R^2$ scores suggest that DLinear is superior. However, DLinear also shows a higher standard deviation. This indicates that DLinear effectively captures linearity in sequential data, leading to accurate predictions for IoT points with strong linear relationships. Conversely, it struggles with complex inherent dependencies, resulting in poorer performance on datasets with such characteristics.

The overall scores indicate significant potential for improvement. Considering the comprehensive metadata scope provided by the BTS dataset, future work can leverage knowledge graphs to enhance data modality. This approach could improve the accuracy and robustness of deep learning models in zero-shot forecasting.

 https://doi.org/10.52202/079017-4232