# Highlights

**Reproducibility of Machine Learning-Based Fault Detection and Diagnosis for HVAC Systems in Buildings: An Empirical Study**

Adil Mukhtar, Michael Hadwiger, Franz Wotawa, Gerald Schweiger

- The literature is not properly document to support reproducibility.
- Under reporting of resources such as coding scripts and datasets.
- Findings suggest a need for intervention to improve current research practices.

# Reproducibility of Machine Learning-Based Fault Detection and Diagnosis for HVAC Systems in Buildings: An Empirical Study

Adil Mukhtar[a], Michael Hadwiger[a], Franz Wotawa[b], Gerald Schweiger[a]

[a]*Vienna University of Technology, Vienna, Austria*
[b]*Graz University of Technology, Graz, Austria*

**Abstract**

Clear and well-documented research experiments are essential, as they allow researchers to reproduce results and verify the authors' claims. Reproducibility is an inherent characteristic of research and should be given heightened importance, however, the rise of machine learning (ML) techniques has introduced new challenges to achieving it. Thus, both the experimental configuration and the proposed methodology should be documented in a clear and standardized manner. Recently, the use of ML techniques has grown common in building systems scientific disciplines. As this discipline expands in its scope and complexity of its methods, concerns have arisen against this backdrop, much like in other fields, about the materials shared along with the proper documentation to support reproducibility. Therefore, in this study, we aim to analyze ML-based fault detection and diagnosis techniques for building system (BS) and quantify them across a list of reproducibility dimensions. Our analysis revealed concerning findings, indicating that nearly all articles are technically irreproducible due to insufficient disclosure across reproducibility dimensions, i.e., *data*, *preprocessing*, *evaluation*, and especially *hyperparameter* and *code availability*. While reproducibility is becoming increasingly important across scientific disciplines, our author profile analysis shows that most authors in BS have engineering and computer science backgrounds. Applying ML techniques to BS therefore requires targeted interventions to improve reproducibility. These findings highlight the need for intervention and reproducibility guidelines. With this, we aim to contribute to the development of more transparent, reliable, and verifiable research in the field.

*Keywords:* Fault Detection and Diagnosis (FDD), Energy Systems, Machine Learning (ML), Reproducibility, Transparency, Methodological Review, Open Science

## 1. Introduction

Buildings account for almost 40% of global energy use and contribute approximately 20% to global carbon emissions (Schweiger et al., 2020). This significant share underscores the critical role of the building sector in environmental sustainability and in achieving global energy efficiency. The timely advancement of, perhaps serendipitously aligned with growing environmental concerns, Internet of Things (IoT) technologies has not only guided the evolution of building development towards the concept of *smart buildings* (Snoonian, 2003) but also supported and accelerated their growth (Jia et al., 2019). Smart buildings, in the context of Information and Communication Technology (ICT), are defined as cyber-physical systems that enable intelligent decision-making through continuous monitoring and control of building operations. This is achieved through real-time data acquisition and communication facilitated by an IoT layer integrated into the building system architecture. Like any other complex system, building systems are prone to faults, which can lead to undesirable outcomes such as increased energy waste, compromised occupant comfort, and high maintenance costs.

Over the years, many approaches have been proposed to detect defective states and identify potential causes in various components, primarily Heating, Ventilation, and Air Conditioning (HVAC) systems. These methods are commonly referred to as fault detection and diagnosis (FDD) or, synonymously, automated fault detection and diagnosis (AFDD) (Katipamula and Brambley, 2005a,b; Chen et al., 2022). In the context of smart buildings, for example, faults can arise from various source such as sensor failures, control errors, or equipment degradation each leading to distinct performance issues. A fault in an HVAC system can result in excessive energy consumption, reduced thermal comfort, or even system failure. Depending on the type and severity of the fault, different detection and diagnostic strategies may be required. Early detection and diagnosis of faults is crucial, as studies have shown that operational faults account for approximately 15–30 % of energy losses in commercial buildings (Nelson and Culp, 2022).

Among the proposed FDD methodologies, some rely on *model-based* approaches using mathematical representations of system dynamics, while others adopt *knowledge-based* strategies guided by expert rules and predefined fault signatures (Katipamula and Brambley, 2005a,b). However, more recently, *machine learning-based* approaches have gained

*Email addresses:* adil.mukhtar@tuwien.ac.at (Adil Mukhtar), michael.hadwiger@tuwien.ac.at (Michael Hadwiger), wotawa@tugraz.at (Franz Wotawa), gerald.schweiger@tuwien.ac.at (Gerald Schweiger)

importance and became popular. These methods employ machine learning and statistical models to identify faults by analyzing large datasets from building automation systems (BAS). Proposals in this category range from traditional supervised learning techniques to advanced deep learning architectures capable of detecting subtle anomalies in high-dimensional space (Matetić et al., 2023; Chen et al., 2023b). While numerous studies have reviewed FDD methods for building systems, a summarized classification is provided in Figure 1. In the following, we briefly describe the motivation for conducting this study, along with the objectives and contributions of the study.

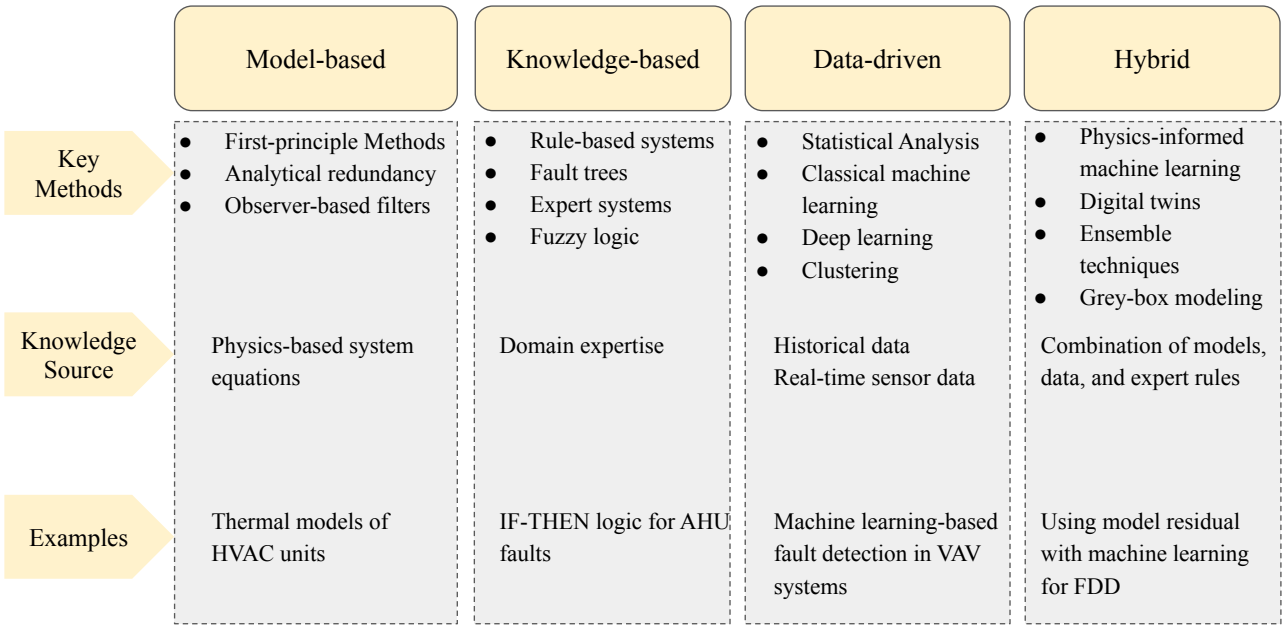


Figure 1: Overview of categorized FDD methods for building systems

### *1.1. Motivation and Background*

Machine learning approaches have gained significant prominence in the field of building systems (Matetić et al., 2023; Chen et al., 2023b; Mirnaghi and Haghighat, 2020). The terms *data-driven* and *machine learning* are often used interchangeably in the literature. In this article, we refer to FDD methods as data-driven if they perform fault detection and/or diagnosis by training a model on data or by employing statistical analysis techniques (Chen et al., 2023b). Nevertheless, these methods utilize historical operational data to learn patterns indicative of both normal and faulty behavior, employing algorithms such as artificial neural networks (Jones, 2015), support vector machines (Namburu et al., 2007), Bayesian networks (Xiao et al., 2014), and various clustering techniques (Du et al., 2014; Li and Hu, 2018; Capozzoli et al., 2015). More recently, researchers have investigated the application of advanced language models (e.g., GPT-3.5/4) for FDD and other energy services in buildings, highlighting both limitations and potential (Zhang et al., 2025; Liu et al., 2025; Langer et al., 2025).

Despite the growing adoption of data-driven methodologies and the increasing sophistication of machine learning models in recent research, we observed sustained shortcomings in the transparency of model development and evaluation procedures. For instance, Haibe-Kains et al. (2020) highlighted even though McKinney et al. (2020) compellingly demonstrated the potential of AI techniques in advancing medical imaging, its failure to provide sufficiently detailed methodologies, complete coding artifacts, and openly accessible data repositories severely compromises its scientific credibility and erodes trust in its findings, despite the apparent promise of its results. Analogously, Popper (2005) expressed this clearly: "Non-reproducible single occurrences are of no significance for science". Reproducibility enables independent verification of results, the identification of errors, and the cumulative advancement of knowledge. Without it, research findings become epistemologically unreliable and ultimately call into question the progress of the entire field. In practice, this can undermine trust in scientific work — especially in experimental results — and prevent potentially valuable methods, such as those from the field of machine learning, from being applied (Herrmann et al., 2024).

In general, *reproducibility* refers to the ability to independently validate a study's proposed models and claims using the information, documentation, and supplementary materials provided in the manuscript or through externally accessible resources (Gundersen and Kjensmo, 2018). Hence, researchers are encouraged to share their research artifacts, including datasets, executable code, and detailed documentation, alongside clear descriptions of evaluation metrics, experimental procedures, and data partitioning strategies (e.g., train-test splits). Such transparency is essential for ensuring reproducibility and guarding against the risk of "phantom progress" (Ferrari Dacrema et al., 2019), a term used to describe cases where models are insufficiently trained, evaluated on weak baselines, or presented with incomplete experimental details, thereby undermining independent validation and meaningful comparison.

### 1.2. Terminologies & Definitions

A long-standing debate over the terms *reproducibility* and *replicability* across scientific disciplines has often led to greater confusion than a concrete resolution (Drummond, 2009; Reproducibility and replicability committee, 2019). In efforts to distinguish between the two, researchers have proposed definitions grounded not in dictionary semantics, but in the practical and methodological requirements needed to verify the results claimed by a given study (Drummond, 2009; Patil et al., 2016; Reproducibility and replicability committee, 2019). However, for the sake of clarity and to avoid engaging in debates over the nuanced distinction between the two terms, we adopt the definitions provided by (Reproducibility and replicability committee, 2019). This allows us and the reader to maintain consistency and avoid ambiguity throughout this article. According to (Reproducibility and replicability committee, 2019), the terms *reproducibility* and *replicability* are defined as follows:

"*Reproducibility* is obtaining consistent results using the same input data; computation steps, methods, and code; and conditions of analysis."

"*Replicability* is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data."

We also consider the domain specific definition for machine learning presented by Gundersen and Kjensmo (2018) relevant and incorporate through out this review:

"Reproducibility in empirical AI research is the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team."

This distinction is also supported by Nichols et al. (2021). However, in contrast to this view, Drummond (2009) argues that what is commonly referred to as "replicability" in the machine learning community—namely, the exact repetition of experiments through shared code and artifacts—should not be conflated with scientific reproducibility. According to Drummond, scientific reproducibility involves obtaining consistent results through different experiments and under varying conditions, thereby reflecting a more robust validation of scientific claims. His perspective, in our view, aligns reproducibility with the broader concept of methodological generalizability.

Despite these differing interpretations, our primary aim in presenting these definitions is to highlight the ongoing ambiguity surrounding the use of these terms across disciplines. For the purpose of this study, we adopt the definitions established by Reproducibility and replicability committee (2019) and structure our analysis accordingly.

### 1.3. Objectives and Contribution

In this study, we investigate the state of reproducibility within the domain of building systems, with a particular focus on machine learning-based fault detection and diagnosis (FDD) in HVAC systems. To this purpose, we assess whether the field exhibits the same persistent reproducibility issues observed in other disciplines (Haibe-Kains et al., 2020), such as insufficient artifact sharing and information on evaluation procedures, and unavailability of code utility packages. To systematically structure our investigation, we adopt and extend the reproducibility spectrum framework (Peng, 2011)—ranging from basic methodological transparency in publications to full computational replication.

We extend this framework to assess not only externally shared materials, but also the extent to which reproducibility-relevant information is documented within the manuscript itself. This adaptation addresses field-specific constraints while maintaining transparency standards: 1. Many building energy studies face legitimate constraints on data/code sharing (e.g., proprietary systems) 2. Manuscript transparency remains fundamental even when artifacts are available. Hence, we formulate three primary research questions outlined below:

**Primary Research Questions**

**RQ1** : To what extent do data-driven FDD studies for HVAC systems provide complete methodological information such as datasets description, data pre-processing, hyperparameter tuning and model training within their manuscripts?

**RQ2** : What percentage of studies make key artifacts (code, datasets, trained models) available through external sources (e.g., GitHub, institutional repositories)?

**RQ3** : How consistently do studies report train-test splits and evaluation metrics, and are these details verifiable through shared materials?

We queried three major research databases, IEEE Xplore, ACM Digigtal Libraray, and Scopus, to retrieve peer-reviewed conference articles published between 2014 and 2024. We specifically focused on studies addressing FDD in HVAC systems. The initial phase involved a broad retrieval of articles, followed by the categorization of FDD methodologies according to their type. From an initial pool of 327 unique articles, we applied defined inclusion criteria, resulting in 104 eligible articles (≈32% of the original set) for the meta-review. Subsequently, ML-based studies or studies that include at least one ML technique in the proposed FDD approach were identified as the primary studies, in total 66 studies, which were then reviewed in detail with a focus on reproducibility analysis. The detailed review process, reproducibility dimensions, and the resulting findings are presented in subsequent sections of this article. To the best of our knowledge,

this is the first empirical study in the field to conduct a reproducibility analysis of machine learning FDD methods for HVAC systems.

In the following, Section 2 reviews related work on reproducibility and existing research on FDD in building systems. Section 3 introduces the reproducibility characteristics and their relevance to this study, followed by a detailed description of the research methodology in Section 4. The findings and results from our empirical analysis are presented in Section 5. We then share insights and recommendations for researchers in Section 6, and conclude this work in Section 7.

## 2. Background & Related Work

### *2.1. Reproducibility: A growing concern in AI and Engineering Research*

Reproducibility grants credibility, rigor, and confidence to findings published in scientific community. It serves as a cornerstone of scientific research by ensuring transparency and enabling validation claims made by the authors. Unfortunately, this fundamental aspect is often treated as an afterthought, despite its critical role in enabling the discovery of new scientific phenomena and the ground breaking technologies through the accumulation and extension of prior findings. However, over the past two decades, it has become widely acknowledged among scientists that many studies are difficult, or in some cases nearly impossible to reproduce accurately.

Back in 2016, over 1,500 scientists participated in a survey (Baker, 2016), and the responses revealed concerning insights: 70% reported being unable to reproduce experiments published by other authors, and more than half admitted to failed attempts at reproducing their own experiments. Gundersen adn Kjensmo Gundersen and Kjensmo (2018) reviewed 400 studies published in highly ranked artificial intelligence (AI) international conferences (IJCAI[1] & AAAI[2]). Their analysis revealed that none of the reviewed studies fully shared the details required to reproduce the experiments. More specifically, only 20% to 30% of the necessary information was typically shared. In another study (Raff, 2019), the authors aimed to quantify reproducibility by manually attempting to implement 255 studies published between 1984 and 2017. They coined the term *independent reproducibility*, referring to reproducing results without using the code provided by the original authors, as releasing code is sometimes insufficient (Drummond, 2009). Their findings indicated a reproducibility rate of approximately 63%, which is significantly higher than a previous study published by Gundersen (Gundersen and Kjensmo, 2018). Their significance testing further revealed that the strongest empirical relationship was associated with the *readability* of the papers. They suggested focusing on clear and detailed communication of implementation details. We adopt a similar principle in this study, emphasizing that properly written, well-documented, and easy-to-follow papers should be a priority when publishing articles. Pham et al. (2020) conducted an extensive experimental study comprising 2,304 identical runs (144 experimental sets with 16 runs each), requiring over 6.5 montsh of GPU time. Well-known datasets [3] were evaluated using six popular deep learning models [4] using three widely used deep learning libraries [5]. The analysis revealed that, even under default identical training conditions without controlling for NI-factors [6], the accuracy gap between the least and most accurate models could reach as high as 10.8%, even after excluding weak models, i.e., those achieving accuracy below 20%. This striking result highlights the challenges inherent in deep learning research, challenges that, unfortunately, are often overlooked during the development and subsequently during the publishing process.

The share of reproducibility studies is comparatively higher in other scientific disciplines than in energy related research. Researchers in the energy sector appear to be progressively acknowledging the challenges of reproducibility. However, to the best of our knowledge, no empirical studies, especially including machine learning techniques, to date have conducted an in-depth analysis of published articles in this domain to assess the state of reproducibility. That said, we did identify a few studies that offer guidelines aimed at addressing reproducibility challenges in energy research by promoting greater openness and methodological robustness (Huebner et al., 2021; Henry et al., 2021; Verticchio et al., 2024; Shekhorkina et al., 2024).

Huebner et al. (2021) outlines several challenges to reproducibility and proposed adoption of the *TReQ* approach-Transparency, Reproducibility, and Quality. They identified key challenges, including domain contextual sensitivity, the high cost and time demands of trials, and resistance from external partners (e.g., utility companies) who may oppose data sharing due to competitive concerns or non-disclouse agreements (NDAs). However, the proposed solutions through the adoption of *TReQ* approach. Specifically, they recommend the preregistration of studies and their plans (PAPs), adherence to reporting guidelines, i.e., standardized indicators specifying which details should be included in published reports and early submission of research findings for community scrutiny through preprints.

The work by Henry et al. (2021) highlights several challenges inherent in energy systems modeling and model intercomparison efforts within electric sector. For instance, many energy systems have parametric and structural complexity, which makes model-based approaches more challenging compared to model-free (machine learning). As a results model

[1] International Joint Conference on Artificial Intelligence
[2] Association for the Advancement of Artificial Intelligence
[3] MINST, CIFAR-10, and CIFAR100
[4] LeNet-1, LeNet-4, LeNet-5, ResNet-38, ResNet-56, and WideResNet-28-10
[5] Tensorflow, CNTK, and Theano
[6] Non determinism-introducing factors, such as shuffling, weight initialization, data augmentation, etc., affecting the training and final model accuracy.

designed to answer similar questions often result in dissimilar outcomes due to diverging input parameters and structural uncertainty. They presented a benchmarking framework using simplified scenarios applied to four open-source models of the U.S. electric sector. Their finding demonstrate that consistency can be improved by identifying specific structural differences and reducing parametric uncertainity. The authors also highlighted, among other issues-such unreported uncertainities and non-unique solutions in optimization problems-the lack of community-wide benchmarking standard, which is critical for reproducibility. To address these challenges, they call on researchers to increase transparency, enabling verification and improving the identification of sources of discrepancy.

More recently, Verticchio et al. (2024) reviewed 105 articles reporting case studies on thermal comfort improvements and conservation. In total, 112 case studies were identified, as some articles included multiple studies. The primary articles reviewed spanned from 2011 to 2022 indexed in Scopus, although the inclusion criteria target publications from 1997 to 2022. The findings revealed that reproducibility is severely effected by the lack of uniform details regarding model choices, makine it difficult for other researchers to replicate the exact steps or assumptions made in a given study. Furthermore, the authors highlighted the need for "tailored comparative tests to verigy the differences among simulation software tools and algorithms", in other words, standardized test suites to understand why different tools may produce dissimilar results when applied to similar problems. Finally, they call for the establishment of FAIR (findable, accessible, interoperable, and reusable) data repositories to build a critical mass of accessible and high-quality information space.

### *2.2. FDD Methods in Energy Systems*

In the energy building systems sector a lot of energy consumption and therefore costs could be saved by optimizing the control of the building and detection and mitigating faults (Melgaard et al. (2022)). The collective term for processes that (automatically) identify abnormal system behavior and determine the affected component or type of fault is Fault Detection and Diagnosis (FDD).

One approach to FDD are so-called rule-based methods where one defines thresholds or if-then logic to detect irregularities. For this kind of methods, extensive knowledge about the systems needs to be available from a domain expert. The approach can only cover a limited amount of faults, is sensitive for parameter tuning and scales poor to different applications/settings due encapsulation of the system specifics. For example Schein et al. (2006) describe a rule-based system that has a pre-defined set of 28 rule and 5 operation modes to detect several different faults.

In model-based approaches, the physical characteristics of the system are modeled using first-principles or gray-box methods by estimating system parameters (Isermann (2005)). Establishing a physical model can be complex and computationally expensive. One way to cope with that is to use knowledge-based approaches. Thereby the amount of in- and outputs and the model complexity is reduced wherever possible together with domain experts. However, as soon as a model is tailored to a specify application by experts, it loses scalability to other applications (Yang et al. (2014)).

In recent years, the amount of available operational data rose significantly. Therefore, data-driven approaches gained popularity. In this approach the idea is to learn/predict the system characteristics and behavior from historical operational data instead of complex modeling. This also enables a better scalability between different buildings or systems since no expert knowledge of the systems is needed. Within data driven approaches, several different methods (or a combination) can be used, some prominent examples are clustering algorithms, principal component analysis (PCA), support vector machines (SVM) and neural networks (NN) (Chen et al. (2023a)). However, not only the method is important also how the parameters of the method are optimized and in which ranges. Furthermore, the operational data need to be preprocessed before training the model, this can include outlier detection, scaling/transformation, feature selection. All these processes can be applied with different settings and extend, therefore for a reproducibility perspective it is important to document them clearly.

Currently, the application of FDD methods in industry is still limited. Most of the applied models still rely on expert knowledge and expert-defined thresholds; however, this will decrease with the use of data-driven methods. In (Heimar Andersen et al. (2024)) barriers for adoption are listed as the lack of standardization of knowledge and tools, as well as the shortage of available datasets and metadata.

## 3. Foundations and Reproducibility Dimensions

Reproducibility can rarely be assessed as a binary outcome through a simple objective function—for instance, by asking: *"Is this study reproducible?"* Unfortunately, there is no straightforward answer to such simplistic question. As it largely depends on the availability and quality of relevant artifacts and documentation presented in the article. The content within the manuscript alone is insufficient unless it is complemented by external resources, such as shared datasets, code repositories, and related materials. Crucially, the sharing of these resources does not, by itself, ensure reproducibility due to the stochastic nature of machine learning methodologies such as data shuffling, weight initialization (Raste et al., 2022), subtle differences in package versions, data leakage (Semmelrock et al., 2025) and more. Therefore, we believe, to address these nuances with fairness and transparency, a systematic literature review is required to quantitatively assess the *degree* of reproducibility of each study based on a set of reproducibility *dimensions* (Gundersen and Kjensmo, 2018) and a corresponding checklist (as we call them *variables*). Thus, in this survey, we build upon the concept of the reproducibility spectrum (*degree*) presented by (Peng, 2011), as shown in Figure 2. Our goal is to identify and evaluate the information

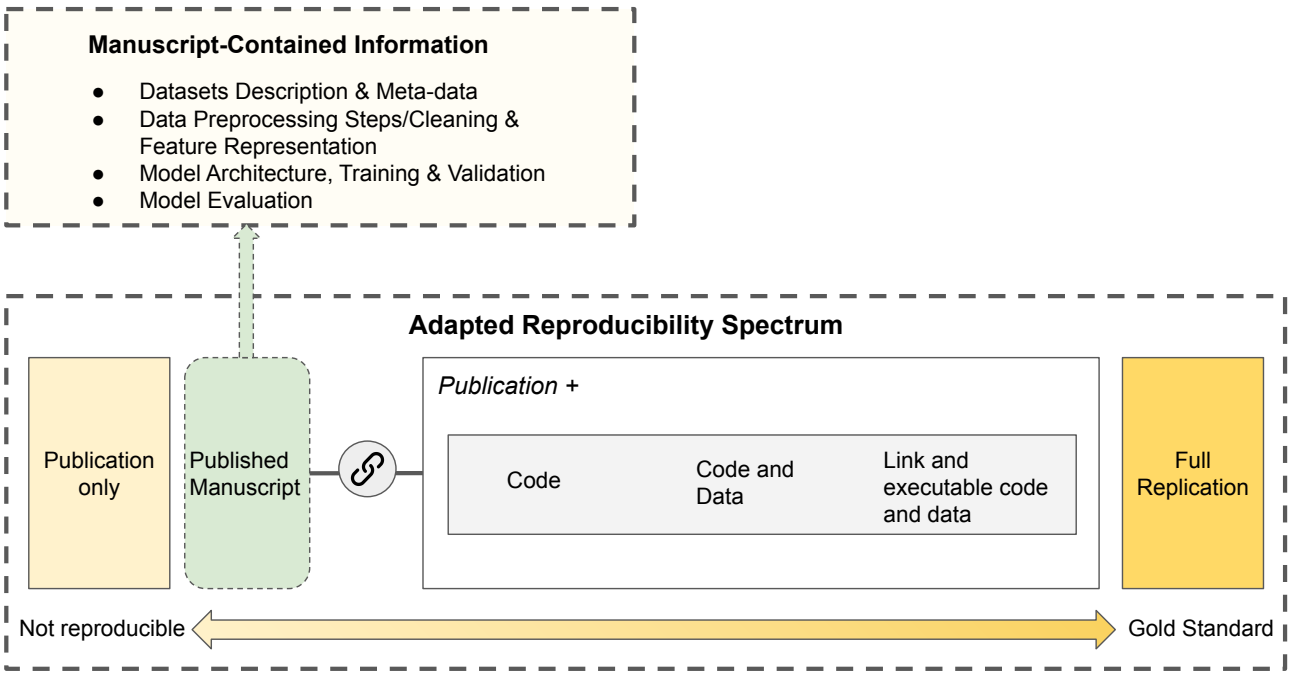


Figure 2: Adapted reproducibility spectrum from (Peng, 2011)

provided within the manuscript that is critical in supporting—and to some extent ensuring—the reproducibility of the study. In particular, we focus on identifying information related to datasets, data preprocessing steps, model development, hyperparameters optimization, evaluation strategies, and accessibility of external artifacts such as executable code, datasets, pre-processing and fine-tuning implementations.

To answer primary research questions, we curated a reproducibility variables checklist adapted from existing reproducibility surveys and guidelines (Gundersen and Kjensmo, 2018; Artrith et al., 2021; Pineau et al., 2021; Mukhtar et al., 2024; Olszewski et al., 2023), specifically for the task of machine learning-based FDD in HVAC energy systems. First, we present a general categorization of the variables into several classes, including dataset, data cleaning/preprocessing, model development, evaluation, and code availability. These variables along with their description and categories are presented in Table 1. The descriptions of these variables are framed as questions to guide the review process, with answers selected based on predefined criteria. For clarity and ease of reference during the discussion of results, we assign each variable a unique identifier. We then group these variables into three reproducibility dimensions as presented by (Gundersen and Kjensmo, 2018): 1. data 2. methodology 3. experiment.

It is important to note that we make a distinction between the categories of variables (Table 1) and the reproducibility dimensions. The categories represent different stages in the development workflow of machine learning-based methodologies. In contrast, the dimensions are an overlapping subset drawn from across categories. These dimensions highlight the critical aspects that jointly influence the reproducibility of a study. As discussed previously, we follow the approach of (Gundersen and Kjensmo, 2018) in categorizing reproducibility into three distinct dimensions. We assess each dimension using the concept of reproducibility degree (Peng, 2011), where a higher degree indicates that the reviewed articles offer more comprehensive information and materials to support reproducibility. While preserving the core intent of these dimensions, we carefully reinterpret and refine their definitions to better align with our study's scope. Since our approach involves thorough manual article review, *without conducting direct experimental replication*, we assess reproducibility variables through this evaluation. Recall that the guiding principle of this study is that all necessary concrete information and materials for reproducibility should be transparently documented, clearly structured, and straightforward to understand. Furthermore, across all dimensions, the studies' objective, i.e., data-driven based FDD in HVAC systems, serves as a foundational requirement. We now define the reproducibility dimensions as follows:

$D_1$ **(Data)**: Datasets, along with their types and descriptions, serve as essential references for reproducing results and conducting experiments.

$D_2$ **(Method)**: The information related to data pre-processing steps, feature engineering techniques, and model development procedures along with the coding scripts is essential for methodological and results reproducibility.

$D_3$ **(Experiment)**: To support reproducible experiments and results, comprehensive documentation and the sharing of code package related to data, optimal methods, evaluation procedures, and metrics are essential.

Table 1: Categorization of Reproducibility Variables

| **Dataset** | | |
|---|---|---|
| **Identifier** | **Variables** | **Answers** |
| $data_{listed}$ | Is the dataset (or datasets) listed? | 1 / 0 |
| $data_{metadata}$ | Are the metadata and description of the dataset (or datasets) provided? | 1 / 0 |
| $data_{stats}$ | Are the relevant statistics discussed, e.g., the number of samples, etc.? | 1 / 0 |
| $data_{type}$ | What is the type of the dataset(s)? | Real-world / Simulation / Experiment / No Information |
| $data_{access}$ | Is information about the accessibility of the dataset(s) shared? | Purchasable / Public / Proprietary / No Information |
| **Data Cleaning/Preprocessing & Feature Representations** | | |
| **Identifier** | **Variables** | **Answers** |
| $preproc_{data}$ | Are the data preprocessing steps documented? | 1 / 0 |
| $preproc_{features}$ | Are data to feature representation methods clearly described? | 1 / 0 |
| $multiple\ data$ | If multiple data sources are used, is their integration clearly stated? | 1 / 0 / No Information |
| **Model Training & Validation** | | |
| **Identifier** | **Variables** | **Answers** |
| $opt_{mentioned}$ | Is hyperparameter optimization mentioned for the proposed model(s)? | 1 / 0 |
| $opt_{baseline}$ | Is hyperparameter optimization mentioned for the baselines? | 1 / 0 |
| $opt_{procedure}$ | Is the hyperparameter optimization procedure described? | 1 / 0 |
| $params_{models}$ | Are hyperparameter search ranges reported for the proposed model? | 1 / 0 |
| $params_{baseline}$ | Are hyperparameter search ranges reported for the baselines? | 1 / 0 |
| $params_{best\ model}$ | Are the best hyperparameters reported for the proposed model? | 1 / 0 |
| $params_{best\ baseline}$ | Are the best hyperparameters reported for the baselines? | 1 / 0 |
| **Evaluation** | | |
| **Identifier** | **Variables** | **Answers** |
| $eval_{splitting}$ | What type of data splitting is reported? | Single split / Train-Test-Validation / Cross Validation |
| $eval_{metrics}$ | Are the metrics used for evaluation reported? | 1 / 0 |
| $eval_{sig\ test}$ | Are the details of statistical significance testing provided? | 1 / 0 |
| **Code Repository** | | |
| **Identifier** | **Variables** | **Answers** |
| $code_{link}$ | Is the link to the code repository available? | 1 / 0 |
| $code_{empty}$ | Is the code repository empty? | 1 / 0 |
| $code_{preproc}$ | Is the source code for data preprocessing provided in the repository? | 1 / 0 |

| | | |
|---|---|---|
| $code_{feature\ gen}$ | Is the source code to generate features of the dataset provided? | 1 / 0 |
| $code_{eval}$ | Is the source code for evaluation provided in the repository? | 1 / 0 |
| $code_{params\ opt}$ | Is the source code for hyperparameter tuning provided in the repository? | 1 / 0 |
| $code_{info}$ | Is supplementary code info (e.g., README, requirements.txt) provided? | 1 / 0 |
| $code_{runnable}$ | Is a runnable model implementation provided (e.g., Docker)? | 1 / 0 |

These dimensions collectively determine the overall degree of reproducibility. For example, a study that provided details on hyperparameter optimization and includes the corresponding code package may still be irreproducible if no information is provided for the dataset used. These dimensions are overlapping subsets derived from the reproducibility variables as shown in Figure 3. The figure presents the list of *variables* and illustrates their association with the three dimensions, including overlaps across them. Dimensions $D_2$ and $D_3$, referring to the method and experiment, respectively, are predicated on the availability of the dataset(s), i.e., $D_1$. The methodologies proposed in scientific articles, especially those based on machine learning, are often presented as frameworks, pipelines, or algorithmic pseudocode. Accordingly, we identified key variables such as data processing, feature generation, hyperparameter optimization, ranges for hyperparameters during the optimization, and the availability of a code repository link, all of which are essential for reproducing the method. This dimension also intersects with Dimension $D_1$, as re-running the methodology requires access to the original datasets. As for dimension $D_3$, we consider it the desirable outcome of reproducibility, as it focuses on replicating the results of the proposed technique, comparing them with established baselines, and validating the experimental outcomes reported in the article. This dimension includes variables related to dataset accessibility, feature generation, data preprocessing, the corresponding coding scripts, evaluation strategies, data splitting techniques, and the best-performing parameters for both the proposed and baseline models.

So far, we have described the reproducibility *variables* (Table 1) and defined the dimensions that span these variables. The rationale behind dissecting the reproducibility variables into three overlapping subsets is to assess the degree of reproducibility, i.e., how well the variables within each dimension are documented. Our next objective is to quantify each proposed study by assigning a reproducibility score, reflecting the extent to which information and materials are shared to support reproducibility. To achieve this, we adapted the quantification scheme from (Gundersen and Kjensmo, 2018). We first evaluate each dimension separately, followed by computing an overall aggregated reproducibility score across all dimensions. For a given study $s$, we first define a function $v_j(s)$ for the $j^{th}$ variable as follows:

$$v_j(s) = \begin{cases} 1, & \text{if study } s \text{ reports variable } j \\ 0, & \text{otherwise OR “No Information”} \end{cases}$$

The dimensions $D_1$ (Data), $D_2$ (Method), and $D_3$ (Experiment) are quantified as follows: for a given dimension $i$ where $i \in \{1, 2, 3\}$

$$D_i(s) = \frac{\sum_{j=1}^{D_i} v_j(s)}{|D_i|} \tag{1}$$

For clarity and brevity, a study $s$ is assigned a score of 1 for a dimension only if all variables within that dimension are documented. In other words, it represents the proportion of variables documented for dimension $i$, ranging from 0 (no documentation) to 1 (full documentation). However, to quantify how many studies fully document all variables across dimensions, i.e., reproducible, we define the degree as follows:

$$degree(s) = \frac{\sum_{j=1}^{V} v_j(s)}{|V|} \tag{2}$$

where $V = D_1 \cup D_2 \cup D_3$. In our study, just like in (Gundersen and Kjensmo, 2018), we treat each variable equally and assign uniform weight, i.e., 1. The reason for treating them as equally important is that the variables are carefully curated to reflect fundamental aspects of research methodology in the building system scientific discipline for the ML-based FDD. For example, one of the key barriers to open science in building system field is the use of proprietary datasets, or datasets that cannot be made openly available due to non-disclosure agreements (NDAs) and strict data-sharing policies and regulations. To address this, we assign a score of 0 only when no information about such details is provided. However, in cases where the dataset is, for example, purchasable, we still consider this crucial information and record 1—provided that it is explicitly mentioned in the study and includes a reference or external link. Such documentation enables independent research teams to access the dataset if needed.

We conclude this section by emphasizing that information in a manuscript should be clearly reported and easy to find. Clear documentation is essential for ensuring the reproducibility of proposed studies, while ambiguity challenges the ability to achieve it. To address this, we carefully designed a list of reproducibility variables, organized them according to key

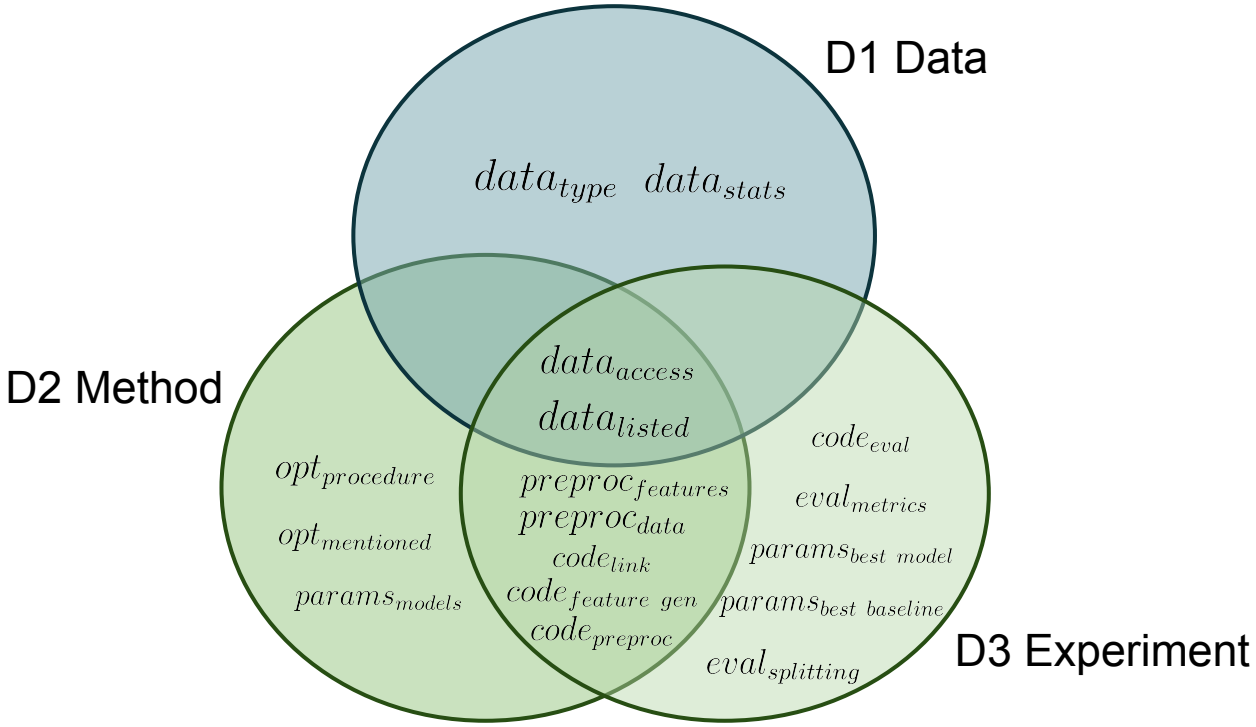


Figure 3: Reproducibility Dimensions

aspects of research methodology, and quantified them to provide insights into the current state of reproducibility in building system for machine learning-based FDD in HVAC systems. In the next section, we detail our survey methodology, including the retrieval of articles and the manual review process.

## 4. Research Methodology

This study adopts a comprehensive systematic literature review (SLR) (Budgen and Brereton, 2006) to guide the articles review process focused on the application of machine learning-based FDD for HVAC systems in building system. The review process is structured into three main phases: 1. planning the review 2. conducting the review, and 3. reporting the review. Figure 4 provides a bird's-eye view of the research methodology used in this study.

In the first phase, *planning*, research questions are defined (see Section 1.3) alongside the development of a search protocol. This includes identifying relevant keywords and selecting scientific databases. Once the initial pool of articles is retrieved by querying these databases using defined inclusion and exclusion criteria, a preliminary categorization is conducted to filter articles that fit the study's objectives and research questions. A meta-analysis of this categorization is presented later in Section 5.1. In the second phase, *conducting*, the manual review, by two researchers, is carried out in three rounds. The details of this phase are provided in Section 4.3. Finally, in the third phase, *reporting*, the results and key insights are presented and discussed.

### *4.1. Databases & Search Strategy*

To ensure comprehensive coverage, the following scientific databases are used to retrieve relevant articles in this study: IEEE Xplore, ACM Digital Library, and Scopus. The main reason for this selection is their extensive collections of scientific publications spanning a wide range of disciplines, including computer science, engineering, artificial intelligence, and other multidisciplinary fields.

Before designing the search query to retrieve articles from the selected databases we consulted with the domain experts in building system to identify and shortlist relevant keywords related to the domain. This preliminary step is taken to warrant a certain level of confidence that the retrieved results, i.e., articles, are relevant and aligned with the scope and objectives of this study. The list of identified key terms is presented in Table 2. Based on this shortlisting, the final search query is constructed using a combination of logical operators, mainly conjuctions and disjunctions, and then used to query the databases. The complete query is presented below:

(fault detection and diagnosis OR fdd OR fault detection and diagnosis) AND
(hvac OR heating ventilation and air conditioning OR chillers OR ahu OR air handling unit OR vav OR variable air volume OR air conditioning) AND
(building energy systems OR building systems OR energy building OR buildings)

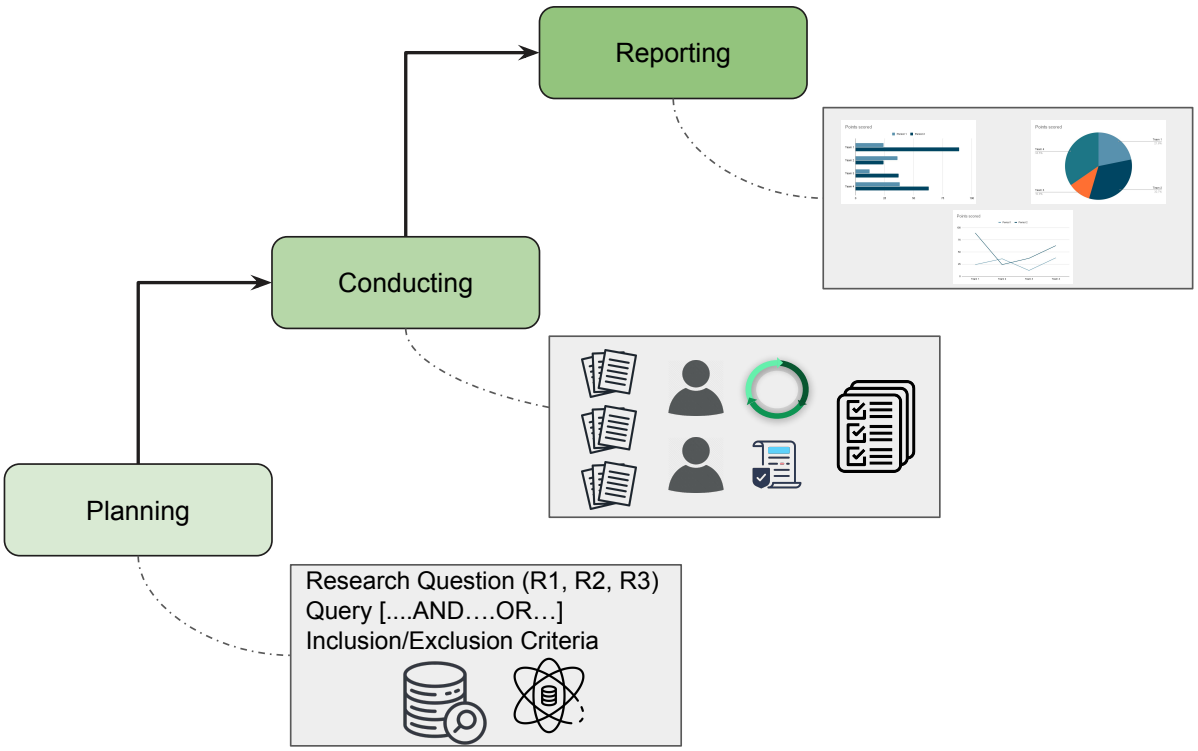


Figure 4: Research Methodology

| | | | |
|---|---|---|---|
| fault detection and diagnosis | fdd | fault detection | fault diagnosis |
| hvac | heating ventilation and air conditioning | chillers | ahu |
| air handling unit | vav | variable air volume | air conditioning |
| building energy systems | building systems | energy building | buildings |

Table 2: List of key terms identified

### *4.2. Articles Selection Criteria*

Querying scientific databases often results in many false positives, even when using relevant keywords, and may also return outdated or irrelevant publications. For example, in our case, we are specifically interested in reviewing articles published in conferences within a defined date range. Therefore, it is important to establish clear and well-defined selection criteria to ensure that the results remain aligned with the predetermined objectives and goals of the study. More specifically, we define our inclusion criteria such that if a paper meets these conditions then it is selected for review, consequently, its negation serves as the exclusion criteria and must be false. The inclusion criteria are as follows: The article

1. is related to energy building systems.
2. proposes a methodology for heating ventilation and air conditioning systems.
3. presents an application for fault detection and diagnosis in HVAC systems.
4. is published in the conference proceedings.
5. is published in 2014 or later.
6. is written in English.
7. is not a survey paper.

Once the search strategy, database selection and selection criteria are established, an initial set of articles are retrieved. The search query resulted in a total of 368 articles published in conferences related to building systems, engineering, artificial intelligence, computer science, and their interdisciplinary fields. These articles cover a range methodological approaches, including model-based, rule-based, knowledge-based and machine learning-based techniques.

### 4.3. Articles Extraction & Reviewing Process

The overall steps involved in the *planning* and *conducting* phases, along with their outcomes, are presented in Figure 5. The query resulted in a total of 327 unique articles out of 367 retrieved. In our experience with Scopus, similarly to Verticchio et al. (2024), the search results included a high number of false positives relative to the provided keywords. This significantly increased the reviewing effort. In contrast, the articles retrieved from ACM and IEEE Xplore were generally more relevant with higher precision compared to those from Scopus. Nevertheless, the first manual review focused on classifying these articles into one of the following categories: model-based (MB), knowledge-based (KB), machine learning-based (ML), or a combination of these approaches. This classification process resulted in 104 relevant articles. From these, 66 were selected as primary articles each including at least one ML technique in the proposed methodology for fault detection and diagnosis (FDD) in HVAC systems[7].

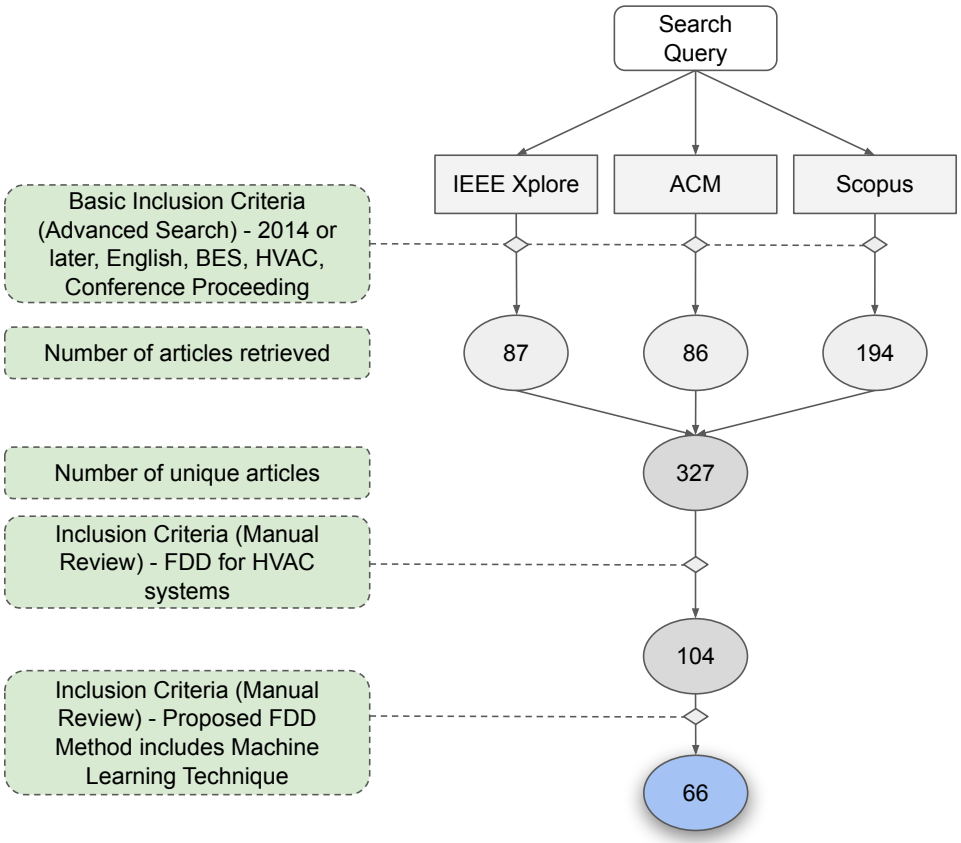


Figure 5: Articles Retrieval Process Flowchart

Two authors independently reviewed the primary articles using the predefined list of reproducibility variables. As briefly mentioned earlier, this phase is completed in three rounds. In the first round, both authors reviewed a small subset of articles to establish consistent understanding of the reproducibility variables and to identify any ambiguous parts requiring clarification or refinement. In the second round, all relevant articles were reviewed independently. In the third round, the outcomes for each variable and article were compared, and any discrepancies were resolved through discussion. In general, conflicts observed during these rounds were minimal and resolved after discussions. This is mainly because of the preliminary round, which led to an improved understanding of the variables, their definitions, and the overall review process.

## 5. Results & Findings

This section begins with an overview of the methodologies employed in the past decade (Section 5.1). We then present our descriptive and qualitative analysis of the reproducibility variables to address the primary research questions (Section 5.2). This is followed by the quantification of the reproducibility dimensions and an overall assessment of articles in terms of degree of sharing crucial information is presented in Section 5.3.

[7]The reviewed articles, along with the recorded reproducibility variables, are available here: https://github.com/tuw-isab/reproducibility-analysis-ml-based-fdd-hvac

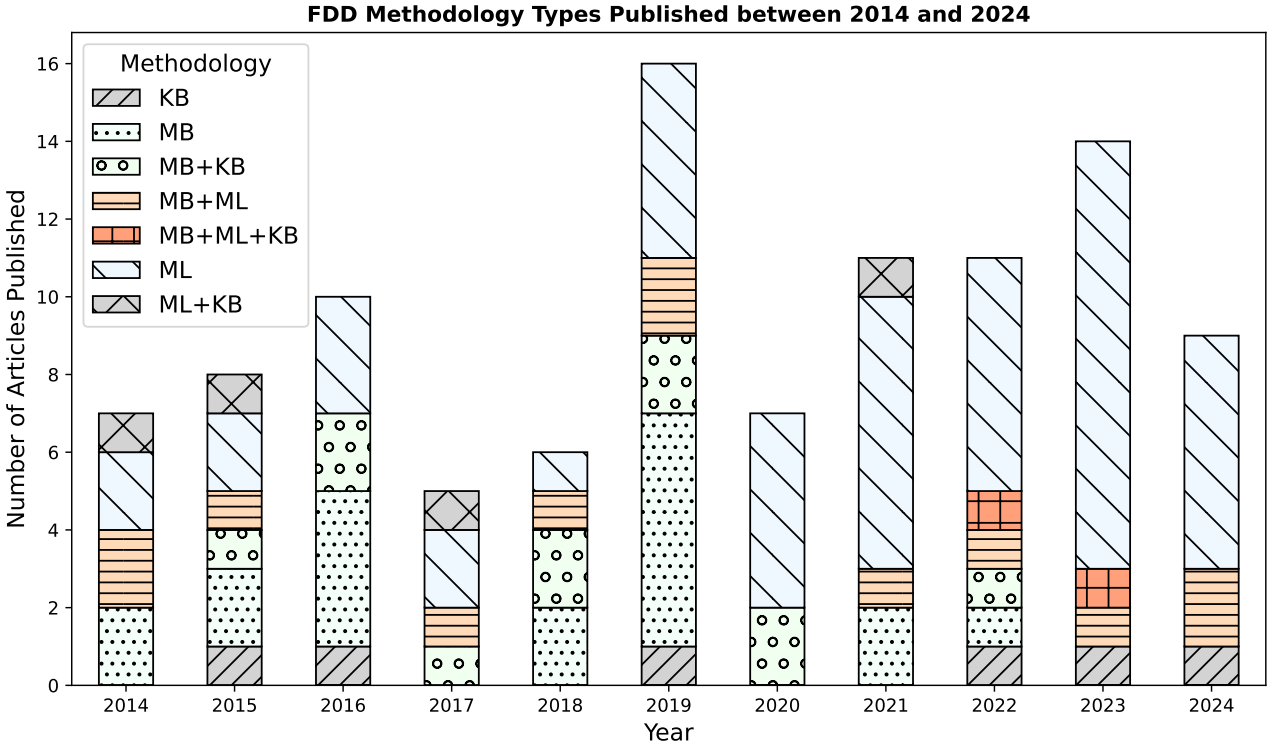


Figure 6: Number of articles and types of FDD methodologies published (2014-2024)

### *5.1. Overview of Fault Detection and Diagnosis Methodologies*

Figure 6 provides an overview of the articles published between 2014 and 2024, specifically addressing FDD in HVAC systems. A key observation is the general increase in publications after the initial five years. This increase can be attributed to the growing adoption of Internet of Things (IoT) technologies, such as smart-meters, real-time monitoring, and building automation systems, which have matured significantly over the past decade (Moudgil et al., 2023; Ahmad and Zhang, 2021). According to (Ahmad and Zhang, 2021), the number of connected IoT devices is projected to reach 75.4 billion by the end of 2025, nearly five times the number back in 2015. More importantly, numerous studies have consistently demonstrated that IoT integration in building systems contributes to substantial improvements in energy efficiency, cost reduction, and environmental sustainability; some examples are (Rohayani et al., 2024; Poyyamozhi et al., 2024; Moudgil et al., 2023).

A closer examination of the publication trends reveals that machine learning (ML)-based approaches have been consistently represented throughout the review period, exhibiting a clear upward trajectory. Standalone ML-based FDD studies account for the largest share, comprising 48% of the publications; this share further increases up to 65% when considering hybrid approaches that combine ML with model-based and knowledge-based techniques. This growth underscores the increasing reliance on ML techniques for FDD in HVAC systems. We believe that this increase is driven by the growing recognition of the effectiveness and benefits of ML techniques across a range of research domains, including software engineering (Abid et al., 2021; Aleem et al., 2015; Mukhtar et al., 2022), healthcare (Maity and Das, 2017; Sharma et al., 2014), and agriculture (Tripathi and Maktedar, 2016; Lu et al., 2017; Amara et al., 2017), among others. In contrast, model-based methods (18%) and hybrid approaches (22%), those combining model-based (MB) with either knowledge-based (KB) or machine learning components, also appear regularly in the literature but do not exhibit the same degree of growth. While these methods remain relevant, their adoption trajectory is comparatively less pronounced than that of purely ML-based techniques. Techniques that integrate all three components, model-based, machine learning-based, and knowledge-based (MB+ML+KB), constitute only a minor portion of the literature, representing approximately 2% of the reviewed studies. The remaining variants, including standalone KB methods and their hybrid combinations (e.g., ML+KB), collectively account for only approximately 10% of the reviewed studies.

Publications in recent years clearly reflect a shift in methodological focus, with ML approaches increasingly surpassing model-based and knowledge-based reasoning methods in terms of research activity. In the following subsections, we turn our attention to answer the primary research questions and presenting our findings and insights related to reproducibility aspects. To this end, we systematically reviewed a subset of articles that propose FDD techniques for HVAC systems and incorporate at least one ML technique. This includes studies classified under the ML category as well as those employing hybrid approaches that combine ML with model-based or knowledge-based methods.

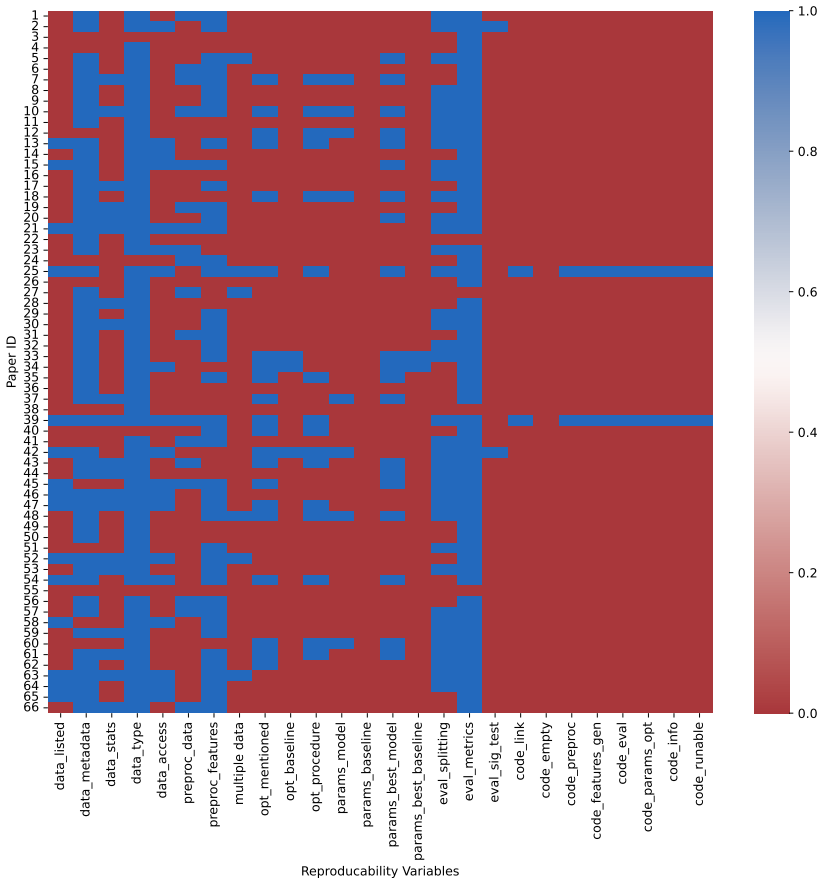


Figure 7: Heatmap of the Reproducibility Variables

### *5.2. Reproducibility Variables Analysis*

We start with the information recorded for the list of variables presented in Table 1. Figure 7 presents, in the form of a heatmap, an overall assessment of the research articles that have documented reproducibility-related variables. It is evident that none of the variables are fully documented across the reviewed studies. In particular, information sharing related to model training and code packages appears to be among the least reported categories. This is even more pronounced in the case of code package sharing, with only a few exceptions. We now share results and analysis to answer the primary research questions.

**RQ1**

The literature is not well-documented, and many of the variables are not fully reported across the reviewed articles.

Figure 8 shows the extent to which information is documented for variables related to dataset characteristics, preprocessing, model development and its fine tuning. It can be seen that not all variables are reported, revealing significant information gaps. For the dataset characteristics (Figure 8(a)), the dataset used in the study is reported in only 23% of the articles, while its basic statistical details such as the number of samples, mean, portion of missing data, are reported in just 30% of the articles. These findings indicate that the preliminary requirements for reproducibility are largely missing in the reviewed articles. Despite these shortcomings, a noteable 80% of the articles provided meta-data about the datasets. This inlcudes information related to environmental context and data collection details such as time duration, building size, number of floors, number of zones, and, in the case of experimental datasets, descriptions of the laboratory setup and other relevant information. Interestingly, this outcome aligns with findings by (Gundersen and Kjensmo, 2018), where about 49% of articles reported data-related information—the most frequently documented variable. However, our formulation is more granular, while theirs abstract, framed as: *"How well is the dataset documented?"*. In our case, we believe the reason behind the relatively high level of information sharing of this variable, compare to others, lies in the primary research background of the authors. Researchers and engineers in the building energy domain often focus on development model- heuristics- and knoweldge-based techniques, which require an in-depth understanding of building systems and their operational environments. As a result, they tend to inherently provide more comprehensive metadata in such cases.

A closer examination combined with the analysis of data type and accessibility (see Figure 9(b)) reveals that approximately 71% of the articles do not provide any information regarding the nature of the dataset, i.e., whether it is public, proprietary, or commercially available. However, the majority articles utilize real-world datasets, followed by simulated

data. This is a concerning outcome, as one might reasonably assume that real-world datasets are often not made publicly available due to NDAs and data sharing regulations. However, the majority of the articles fail to mention such constraints in their description of dataset accessibility. Alternatively, the absence of any statement regarding dataset accessibility might itself be interpreted as an implicit indication of proprietary constraints—although, in the absence of clarification, one can only assume. Furthermore, we found that nearly half of the articles reported training and testing their methodologies on real-world datasets. This may also reflect the presence of contractual agreements or NDAs with organizations or companies, which may have restricted authors from disclosing detailed information. Nevertheless, it is notable that more than half of the articles make no mention of dataset accessibility or related guidelines. Reliance on simulated datasets is reported in 28% of the articles, followed by the combined use of both real-world and simulation data in 12%. This type of setting is commonly observed in studies where the methodology is trained on simulated data and tested on real-world datasets. Finally, 7% of the articles report using experimental datasets, while a similar proportion provide no information regarding the type of dataset used.

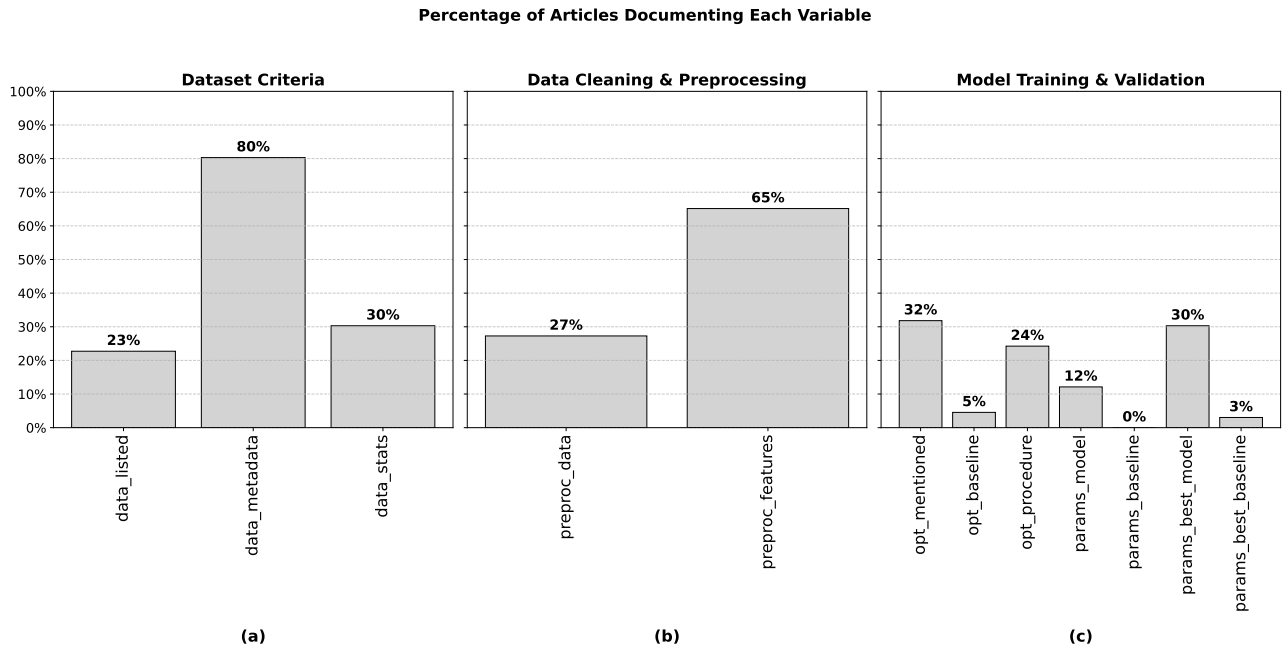


Figure 8: Percentages of Variables Recorded

Looking at Figure 8(b), data preprocessing related information is reported in only 27% of the articles, whereas feature generation and their representation are documented in 65%. We observed consistent use of basic feature preprocessing techniques such as MinMax Scaler and Standard Scaler. For feature representation, we found that articles typically provided input features along with their descriptions, either in tabular form or within the methodology section. Nevertheless, the disparity between these two variables highlights a concerning trend: although many studies emphasize how raw data is transformed into feature represesntation and input vectors, the initial and equally important part of cleaning and preparing raw data is often under reported.

Figure 8(c) provides insights into the model training and validation phase, one of the most critical steps in the machine learning model development process. Unfortunately, despite its importance, this category scores the lowest on average among all reproducibility aspects addressed in RQ1. While 32% of the reviewed articles mention the use of optimization ($opt_{mentioned}$) only 24% explain the optimization procedure ($opt_{procedure}$), and a mere 5% report how the baseline model was optimized ($opt_{baseline}$). Even more concerning, only 12% of articles document model hyperparameters ($params_{model}$), 30% specify the parameters of the best-performing model ($params_{best\ model}$), and just 3% provide the corresponding information for the best baseline model ($params_{best\ baseline}$). Optimization procedures, such as grid search, random search, or Bayesian optimization, directly influence model performance. Without access to the optimization setup, including the search space, objective function, and validation strategy, it is difficult reproduce claimed results or assess whether improvements over baselines are statistically or practically significant (Ferrari Dacrema et al., 2019). Strikingly, no article reports the parameter settings for the baseline model ($params_{baseline}$), which is critical for fair comparisons. The absence of details on baseline tuning does not help, as poorly tuned baselines can give a misleading impression of the proposed model's effectiveness (Ferrari Dacrema et al., 2019).

**RQ2**

The literature significantly underreports the sharing of resources such as coding scripts, datasets, and trained models via external links.

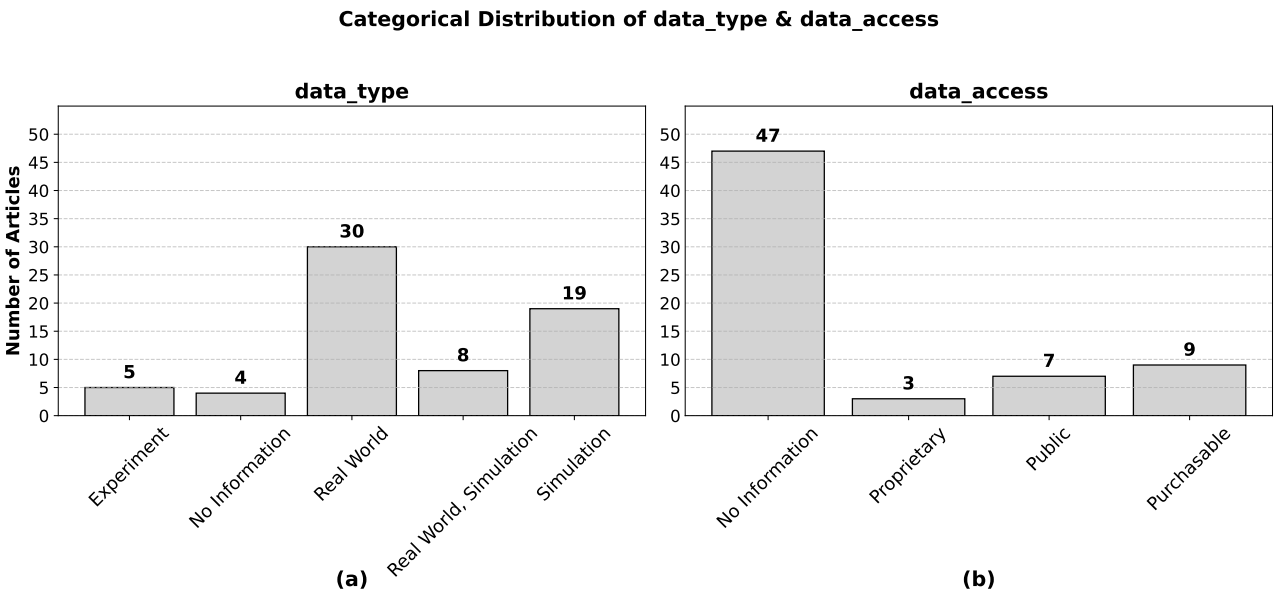


Figure 9: Data Type and Data Accessibility Analysis

Documented information related to code scripts, datasets, and associated executable files is notably scarce. Such resources are typically shared via external links to platforms like Github, Zenodo, or Dropbox. Referring back to Figure 7, here we focus on the variable $code_{link}$, which serves as a predicate for the variables positioned to its right in the figure. In essence, the availability of code is a basic indicator that correlates with the reporting of additional implementation-specific details related to feature generation, preprocessing, evaluation and general information about the code. It is evident that variables in this category are completely unreported across the reviewed articles except for two instances where authors provided link to Github repositories (3%). In one of these cases, the article included a link to Jupyter notebook, but the link was broken and the server returned "Not Found" error. To recover the resource, we performed an additional investigation, identified the author's Github username, and subsequently located the relevant repository.

Positively, we found that both articles shared evaluation code including data splitting strategies, input vectors, comparisons with baseline models, and metrics used to estimate the performance gains. This level of detail also extended, in general, to the search ranges used for model fine-tuning. However, in one case, the code appeared to rely on a predefined set of values for training, whith varying lagged values for input and output sequences. This may have been an oversight, possibly due to the author forgetting the updated or push the final version of the scrip to the repository prior to publication. The baselines consisted of relatively simple models such as Random Forest and Isolation Forest, and in other case where performance on one building was used as a baseline to compare results on another building, an approach called transfer learning, and appeared to be implemented by the authors themselves. However, when authors implement baselines on their own, particularly when comparing their method against results from prior research or state-of-the-art (SOTA) methods, there is a risk that these implementations may not be entirely accurate. This can potentially lead to misleading comparisons and incorrect conclusions (Hidasi and Czapp, 2023). The description in the code repositories ($code_{info}$) is generally limited or minimal. However, we still marked this variable as 1, as some level of information is present.

**RQ3**

Information related to data splitting strategies and evaluation metrics is generally well documented across the reviewed articles.

Overall, in response to RQ3, we found that approximately 59% of the articles reported the type of data splitting ($eval_{splitting}$) technique used, with an even higher proportion of 93% documenting the evaluation metrics ($eval_{mterics}$) used. While data splitting strategies are critical for evaluating model performance, their documentation remains incomplete across many reviewied articles. Looking at Figure 10, which presents the distribution of methodology types, the most frequently reported startegy is single-split validation (33%). This approach, particularly based on random splits, is generally considered suboptimal and not robust against concept drift and other temporal effects (Lyu et al., 2021). However, in our case, the datasets primarily exhibit time-series characteristics where temporal dependencies exist between observation. Even in such settings, it is important to perform evaluation using walk-forward methodologies or time-based cross-validation techniques to ensure that the model's performance is robust, generalizes well to unseen future instances and avoid the influence of NI-factors (Pham et al., 2020). Unfortunately, we found a relatively low rate of reporting cross-validation procedures, only 9% of the articles documented their use. Similarly, 12% of the articles reported using a Train/Test/Validation split, a suitable technique re-adjusting parameter weights during model optimization is necessary.

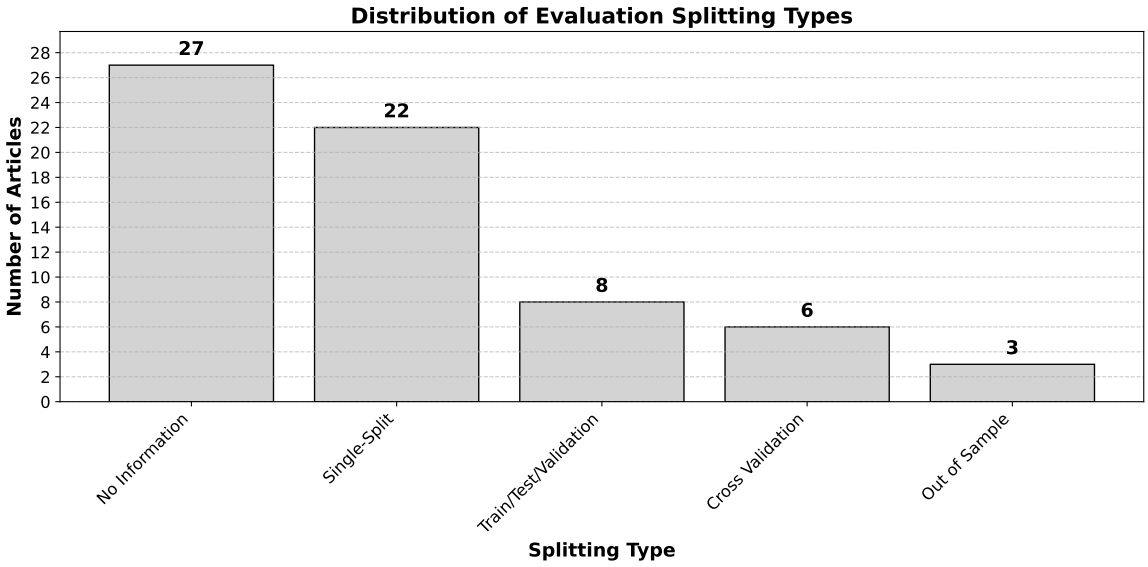


Figure 10: Types of Evaluation Strategies

Such a low rate is also consistent with earlier observations indicating limited reporting of hyperparameter fine-tuning details. Furthermore, only 4% of the articles employed out-of-sample evaluation methods, which involve testing the model on data that was not used during either the training or validation phases. This strategy is particularly useful for assessing model performance on truly unseen data, essentially, its ability to *generalize to an external population* (Varoquaux and Colliot, 2023). However, this technique is not commonly reported in the reviewed articles.

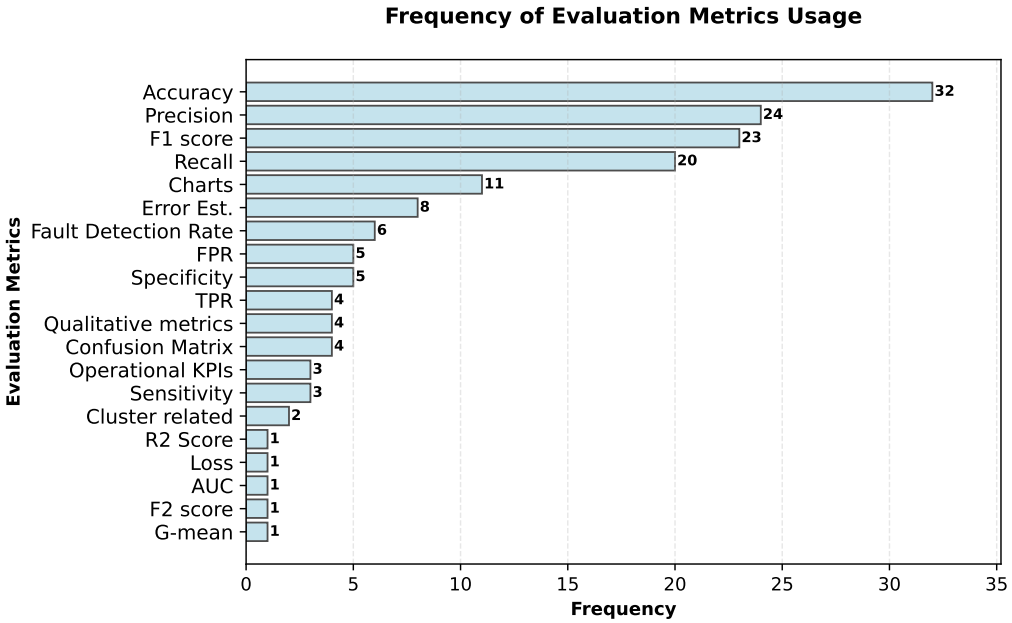


Figure 11: Frequency of Evaluation Metrics

Evaluation metrics are among the most consistently reported indicators in the reviewed articles, with nearly 93% of the studies including them. Figure 11 presents the metrics used, ordered by the frequency of occurenece. A clear cut-off appears after the first four metrics, Accuracy, Precision, F1 Score, and Recall, which are the most prevalent in FDD studies and are consistently reported accort the reviewed literature. In addition, other quantitative metrics such as False Positive Rate (FPR), True Positive Rate (TPR), Error Estimation, Fault Detection Rate, and Confusion Matrix are also commonly used. In some cases, qualitative metrics and operational key performance indicators (KPIs) relevent to building systems are reported, for example, reduction in man-hours, or time elapsed between fault detection and diagnosis or performance monitoring through visualization/charts. In addition, significance testing ($sig_{test}$; see Figure 7) to assess performance gains over appears in only one article. Interestingly, some articles include this test as part of the fault detection methodology; however, this measure typically do not appear in the evaluation phase and remain significantly underreported. Significance testing is common in machine learning scientific community, for instance, the Friedman test often supports comparisons

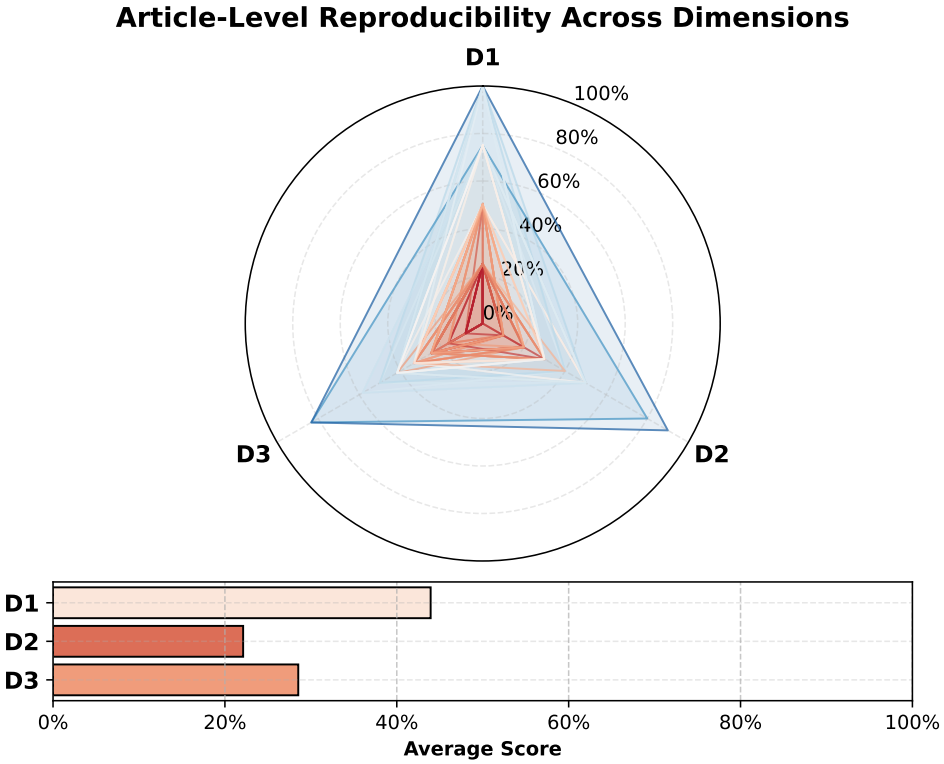


Figure 12: Dimension Scores

of multiple classifiers across multiple datasets (Demšar, 2006).

### *5.3. Assessment of Reproducibility Degree*

In the previous section, we presented detailed insights into the documentation of reproducibility-related variables. The results are particularly concerning for aspects related to model training and the availability of code repositories, where reporting appears limited or, in some cases, significantly underreported. As a next step, we aim to anlayze reproducibility across thee key dimensions: $D_1$ (Data), $D_2$ (Methodology), and $D_3$ (Experiment). This analysis provides a more comprehesive view of how well individual studies support reproducibility crucial components. In this context, Figure 12 presents the quantified assessment of each dimension for the reviewed studies. In most cases, the concentration of measured scores appears near the center of the plot along each axis, on average ranging between 5% and 25%, indicating that the majority of articles fail to report variables more often. A few cases can be easily spotted where the scores are comparatively higher, particularly for dimension $D_1$. In total, six articles document all the variables associated with dimension $D_1$. However, only two articles achieve an overall reproducibility score of approximately 80% for $D_2$ and $D_3$, respectively. This observation may suggest that a few authors inherently adopt more detailed documentation practices, or that certain reporting standards align with requirements imposed by funding agencies or institutional policies.

While the overall scores across all three dimensions (as shown in the subplot) remain relatively low, ranging between 22% and 44%, this indicates that nearly two-thirds of the reproducibility variables are not documented in the reviewed articles, particularly within dimensions $D_2$ and $D_3$. A general trend is that dataset characteristics are reported more frequently; however, this coverage remains insufficient, with over 50% of the variables in this category still undocumented. Although dataset information appears to be the clear winner here, this may reflect the disciplinary background of the authors, many of whom are likely more aligned with building science or industry practice and may not be as engaged with the machine learning research community. Another way to look at it is that $D_1$ includes fewer variables than $D_2$ and $D3$, and some of those, like $data_{metadata}$, are less granular. In contrast, the variables in $D_2$ and $D_3$ have a more detailed and specific to the development of ML methodology, making their absence more noticeable during the review. Further, the assumption underlying this quantification is the independence among the variables. It represents the normalized aggregated sum over the variables for each category. In other words, if the data sources are not listed, the outcome reflects a partial or missing contribution to the overall reproducibility score for the data dimension, regardless of the completeness of other variables in that category.

The overall assessment of reproducibility degree, computed according to Equation 2, is presented in Figure 13. The objective here is to evaluate the extent of information shared across all three dimensions. To this end, the reproducibility variables from dimension $D_1$, $D_2$, and $D_3$ are combined, and a normalized score, referred to as the overall reproducibility degree, is assigned to each article. As shown in the figure, the average reproducibility degree is 31.5%, i.e., on average articles report only about one-third of the information that is crucial to reproduce their experiments. It is important to note

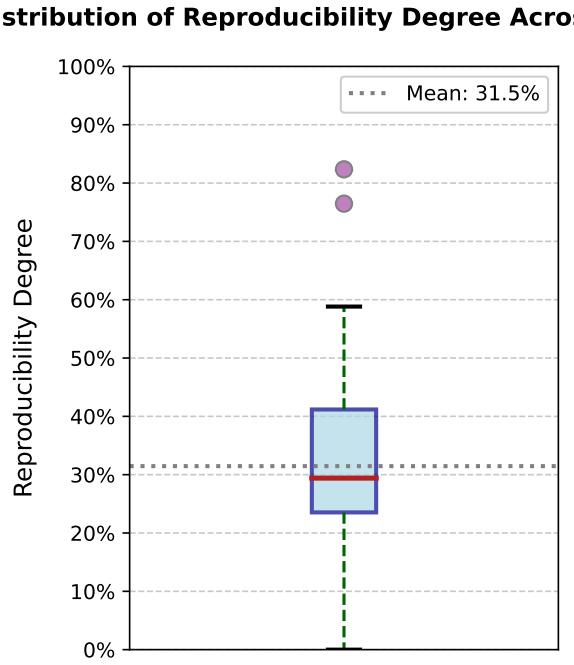


Figure 13: Reproducibility Degree of Articles

that this metric does not require all variables to be reported (i.e., conjunction); otherwise, the score would be zero for all articles. This is a concerning results, as it reflects a broader systemic issue. To put this in perspective, it suggests that articles published over the pas decade offer limited information to support reproducibility, despite growing awareness of its importance in scientific research and the ongoing debate surrounding it (Fidler and Wilcox, 2018; Haibe-Kains et al., 2020; Reproducibility and replicability committee, 2019).

## 6. Discussion

Over the past decade, the term *reproducibility crisis* has become ubiquitous across scientific disciplines, indicating growing concern over the reliability of published results (Baker, 2016; Gundersen and Kjensmo, 2018; Hutson, 2018; Fidler and Wilcox, 2018). Fidler and Wilcox (2018) highlighted reasons for irreproducible results include factors such as the incentive culture of "publish or perish", poorly designed statistical methods, and inadequate publishing practices, all of which are largely common in the behavioral and life sciences. However, one of the primary causes of irreproducibility in the AI/ML scientific field is the insufficient effort put into preparing well-documented articles and comprehensively reporting details on methodology development, testing strategies, and the sharing of external supplementary materials. (Gundersen and Kjensmo, 2018; Hutson, 2018).

We observed an upward trend in publications primarily based on machine learning methodologies. However, the reporting of critical information essential for an independent researcher to replicate the experiments is mostly not available. In particular, we observed almost all of the articles do not share links to external code packages and executeabale scripts. Machine learning techniques are non-deterministic methods by chance, meaning that without explicit control over parameters and clearly defined search spaces, their outcomes may vary due to stochastic elements in the training and evaluation processes (Pham et al., 2020). Therefore, reporting of details related to model development and the search ranges used for hyperparameter tuning is crucial to mitigate the influence of such factors. Unfortunately, we observe that only a small proportion of articles report such details, with an average of approximately 15% providing this information. Building systems are equipped with numerous sensors for monitoring and these measurements are often noisy or incomplete, resulting in information gaps. Additionally, some sensors generate signals only when an event occurs, when there is a change in measurement, or based on predefined heuristics. In such cases, it is important to report statistics related to the datasets, particularly whether the recorded variables (input vectors) contain missing values or exhibit misaligned timestamps. We again observe low rates of reporting for most variables in this category, with the notable exception of information about the facilities, such as experimental laboratories or building structures, which is reported in most articles (80%). We also observed that a large fraction of article do not report the data accessibility type.

Data leakage, i.e., some observations from the testing set are made available during training, is a common issue in the development of machine learning techniques. Various strategies have been proposed to mitigate this problem, particularly in the context of fostering open science and ensuring the validity of reported results (Kapoor and Narayanan, 2023). The likelihood of data leakage becomes even more pronounced when a single-fold split is used and most studies rely on this splitting scheme. It is generally recommended to evaluate a model's performance using the entire dataset

through cross-validation and to analyze the aggregated performance. We also observed the lack of significance testing over baseline techniques to demonstrate the performance gains. Typically, research in machine learning mainly relies on one-to-one comparisons of performance metrics. While such absolute evaluations may show improvements over state-of-the-art methods and baselines, the non-deterministic nature of machine learning models, as previously discussed, can result in lucky outcomes, for example, when the data is biased or not representative of the overall population. Therefore, using reliable evaluation strategies and proper data splitting techniques, along with the analysis of clear and confident improvements over existing methods, is essential to support claims of performance gains.

In general, we observe that none of the reviewed studies can be fully reproduced, as the reporting of information across the main dimensions of reproducibility is significantly lacking. Only two studies achieved higher scores across all dimensions, albeit still insufficient to ensure full reproducibility. The overall reproducibility *degree* is approximately 31.5%, showing that, on average, only about one-third of the necessary information is reported.

## 7. Conclusion

In this work, we have quantitatively demonstrated that research articles published over the past decade on machine learning-based fault detection and diagnosis for HVAC systems are not properly documented to ensure the reproducibility of their results. Our findings suggest a need for intervention to improve current research practices and highlight the importance of establishing guidelines specifically designed to address reproducibility challenges within the scientific discipline of building systems.

## References


Abid, A., Khan, M.T., Iqbal, J., 2021. A review on fault detection and diagnosis techniques: basics and beyond. Artificial Intelligence Review , 3639–3664.

Ahmad, T., Zhang, D., 2021. Using the internet of things in smart energy systems and networks. Sustainable Cities and Society , 102783.

Aleem, S., Capretz, L.F., Ahmed, F., 2015. Benchmarking machine learning technologies for software defect detection. arXiv preprint arXiv:1506.07563 .

Amara, J., Bouaziz, B., Algergawy, A., 2017. A deep learning-based approach for banana leaf diseases classification, in: Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband, Gesellschaft für Informatik eV. pp. 79–88.

Artrith, N., Butler, K.T., Coudert, F.X., Han, S., Isayev, O., Jain, A., Walsh, A., 2021. Best practices in machine learning for chemistry. Nature chemistry , 505–508.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility.

Budgen, D., Brereton, P., 2006. Performing systematic literature reviews in software engineering, in: Proceedings of the 28th international conference on Software engineering, pp. 1051–1052.

Capozzoli, A., Lauro, F., Khan, I., 2015. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Systems with Applications , 4324–4338.

Chen, J., Zhang, L., Li, Y., Shi, Y., Gao, X., Hu, Y., 2022. A review of computing-based automated fault detection and diagnosis of heating, ventilation and air conditioning systems. Renewable and Sustainable Energy Reviews 161, 112395. doi:https://doi.org/10.1016/j.rser.2022.112395.

Chen, Z., O'Neill, Z., Wen, J., Pradhan, O., Yang, T., Lu, X., Lin, G., Miyata, S., Lee, S., Shen, C., Chiosa, R., Piscitelli, M.S., Capozzoli, A., Hengel, F., Kührer, A., Pritoni, M., Liu, W., Clauß, J., Chen, Y., Herr, T., 2023a. A review of data-driven fault detection and diagnostics for building hvac systems. Applied Energy , 121030doi:https://doi.org/10.1016/j.apenergy.2023.121030.

Chen, Z., O'Neill, Z., Wen, J., Pradhan, O., Yang, T., Lu, X., Lin, G., Miyata, S., Lee, S., Shen, C., et al., 2023b. A review of data-driven fault detection and diagnostics for building hvac systems. Applied Energy , 121030.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30. URL: http://jmlr.org/papers/v7/demsar06a.html.

Drummond, C., 2009. Replicability is not reproducibility: nor is it good science, in: Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, National Research Council of Canada Montreal, Canada.

Du, Z., Fan, B., Jin, X., Chi, J., 2014. Fault detection and diagnosis for buildings and hvac systems using combined neural networks and subtractive clustering analysis. Building and Environment , 1–11.

Ferrari Dacrema, M., Cremonesi, P., Jannach, D., 2019. Are we really making much progress? a worrying analysis of recent neural recommendation approaches, in: Proceedings of the 13th ACM conference on recommender systems, pp. 101–109.

Fidler, F., Wilcox, J., 2018. Reproducibility of Scientific Results, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Winter 2018 ed.. Metaphysics Research Lab, Stanford University.

Gundersen, O.E., Kjensmo, S., 2018. State of the Art: Reproducibility in Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence .

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, M.A.Q.C.M.S.B., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., et al., 2020. Transparency and reproducibility in artificial intelligence. Nature , E14–E16.

Heimar Andersen, K., Pommerencke Melgaard, S., Johra, H., Marszal-Pomianowska, A., Lund Jensen, R., Kvols Heiselberg, P., 2024. Barriers and drivers for implementation of automatic fault detection and diagnosis in buildings and HVAC systems: An outlook from industry experts. Energy and Buildings doi:10.1016/j.enbuild.2023.113801.

Henry, C.L., Eshraghi, H., Lugovoy, O., Waite, M.B., DeCarolis, J.F., Farnham, D.J., Ruggles, T.H., Peer, R.A., Wu, Y., de Queiroz, A., et al., 2021. Promoting reproducibility and increased collaboration in electric sector capacity expansion models with community benchmarking and intercomparison efforts. Applied Energy , 117745.

Herrmann, M., Lange, F.J.D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.L., Bischl, B., 2024. Position: Why we must rethink empirical research in machine learning. arXiv preprint arXiv:2405.02200 .

Hidasi, B., Czapp, Á.T., 2023. The effect of third party implementations on reproducibility, in: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 272–282.

Huebner, G.M., Fell, M.J., Watson, N.E., 2021. Improving energy research practices: guidance for transparency, reproducibility and quality. Buildings & Cities , 1–20.

Hutson, M., 2018. Artificial intelligence faces reproducibility crisis.

Isermann, R., 2005. Model-based fault-detection and diagnosis – status and applications. Annual Reviews in Control , 71–85URL: https://www.sciencedirect.com/science/article/pii/S1367578805000052, doi:https://doi.org/10.1016/j.arcontrol.2004.12.002.
Jia, M., Komeily, A., Wang, Y., Srinivasan, R.S., 2019. Adopting internet of things for the development of smart buildings: A review of enabling technologies and applications. Automation in construction , 111–126.
Jones, C.B., 2015. Fault detection and diagnostics of an HVAC sub-system using adaptive resonance theory neural networks. The University of New Mexico.
Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804. doi:https://doi.org/10.1016/j.patter.2023.100804.
Katipamula, S., Brambley, M.R., 2005a. Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i. Hvac&R Research , 3–25doi:10.1080/10789669.2005.10391133.
Katipamula, S., Brambley, M.R., 2005b. Methods for fault detection, diagnostics, and prognostics for building systems—a review, part ii. Hvac&R Research , 169–187.
Langer, G., Hirsch, T., Kern, R., Kohl, T., Schweiger, G., 2025. Large language models for fault detection in buildings' hvac systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 15272 LNCS, 49 – 60. doi:10.1007/978-3-031-74741-0_4.
Li, G., Hu, Y., 2018. Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. Energy and Buildings , 502–515.
Liu, M., Zhang, L., Chen, J., Chen, W.A., Yang, Z., Lo, L.J., Wen, J., O'Neill, Z., 2025. Large language models for building energy applications: Opportunities and challenges, in: Building Simulation, Springer. pp. 1–10.
Lu, Y., Yi, S., Zeng, N., Liu, Y., Zhang, Y., 2017. Identification of rice diseases using deep convolutional neural networks. Neurocomputing , 378–384.
Lyu, Y., Li, H., Sayagh, M., Jiang, Z.M., Hassan, A.E., 2021. An empirical study of the impact of data splitting decisions on the performance of aiops solutions. ACM Transactions on Software Engineering and Methodology (TOSEM) , 1–38.
Maity, N.G., Das, S., 2017. Machine learning for improved diagnosis and prognosis in healthcare, in: 2017 IEEE aerospace conference, IEEE. pp. 1–9.
Matetić, I., Štajduhar, I., Wolf, I., Ljubic, S., 2023. A review of data-driven approaches and techniques for fault detection and diagnosis in hvac systems. Sensors 23. doi:10.3390/s23010001.
McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature , 89–94.
Melgaard, S.P., Andersen, K.H., Marszal-Pomianowska, A., Jensen, R.L., Heiselberg, P.K., 2022. Fault Detection and Diagnosis Encyclopedia for Building Systems: A Systematic Review. Energies URL: https://www.mdpi.com/1996-1073/15/12/4366, doi:10.3390/en15124366.
Mirnaghi, M.S., Haghighat, F., 2020. Fault detection and diagnosis of large-scale hvac systems in buildings using data-driven methods: A comprehensive review. Energy and Buildings , 110492.
Moudgil, V., Hewage, K., Hussain, S.A., Sadiq, R., 2023. Integration of iot in building energy infrastructure: A critical review on challenges and solutions. Renewable and Sustainable Energy Reviews , 113121.
Mukhtar, A., Hofer, B., Jannach, D., Wotawa, F., Schekotihin, K., 2022. Boosting spectrum-based fault localization for spreadsheets with product metrics in a learning approach, in: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, pp. 1–5.
Mukhtar, A., Jannach, D., Wotawa, F., 2024. Investigating reproducibility in deep learning-based software fault prediction, in: 2024 IEEE 24th International Conference on Software Quality, Reliability and Security (QRS), pp. 306–317. doi:10.1109/QRS62785.2024.00038.
Namburu, S.M., Azam, M.S., Luo, J., Choi, K., Pattipati, K.R., 2007. Data-driven modeling, fault diagnosis and optimal sensor selection for hvac chillers. IEEE transactions on automation science and engineering , 469–473.
Nelson, W., Culp, C., 2022. Machine learning methods for automated fault detection and diagnostics in building systems—a review. Energies 15. doi:10.3390/en15155534.
Nichols, J.D., Oli, M.K., Kendall, W.L., Boomer, G.S., 2021. A better approach for dealing with reproducibility and replicability in science. Proceedings of the National Academy of Sciences , e2100769118.
Olszewski, D., Lu, A., Stillman, C., Warren, K., Kitroser, C., Pascual, A., Ukirde, D., Butler, K., Traynor, P., 2023. " get in researchers; we're measuring reproducibility": A reproducibility study of machine learning papers in tier 1 security conferences, in: Proceedings of the 2023 ACM SIGSAC conference on computer and communications security, pp. 3433–3459.
Patil, P., Peng, R.D., Leek, J.T., 2016. A statistical definition for reproducibility and replicability. BioRxiv , 066803.
Peng, R.D., 2011. Reproducible research in computational science. Science , 1226–1227.
Pham, H.V., Qian, S., Wang, J., Lutellier, T., Rosenthal, J., Tan, L., Yu, Y., Nagappan, N., 2020. Problems and opportunities in training deep learning software systems: An analysis of variance, in: Proceedings of the 35th IEEE/ACM international conference on automated software engineering, pp. 771–783.
Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., Larochelle, H., 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). Journal of machine learning research , 1–20.
Popper, K., 2005. The logic of scientific discovery. Routledge.
Poyyamozhi, M., Murugesan, B., Rajamanickam, N., Shorfuzzaman, M., Aboelmagd, Y., 2024. Iot—a promising solution to energy management in smart buildings: A systematic review, applications, barriers, and future scope. Buildings , 3446.
Raff, E., 2019. A step toward quantifying independently reproducible machine learning research. Advances in Neural Information Processing Systems .
Raste, S., Singh, R., Vaughan, J., Nair, V.N., 2022. Quantifying inherent randomness in machine learning algorithms. arXiv preprint arXiv:2206.12353 .
Reproducibility, replicability committee, 2019. Reproducibility and replicability in science. National Academies Press.
Rohayani, H., Ermaini, E., Handayani, R., Rahardian, R., Nanjar, A., 2024. Effect of iot integration in energy management system and grid responsiveness on energy efficiency and cost reduction in jakarta government buildings. West Science Interdisciplinary Studies , 1077–1087.
Schein, J., Bushby, S.T., Castro, N.S., House, J.M., 2006. A rule-based fault detection method for air handling units. Energy and Buildings , 1485–1492URL: https://www.sciencedirect.com/science/article/pii/S0378778806001034, doi:https://doi.org/10.1016/j.enbuild.2006.04.014.
Schweiger, G., Eckerstorfer, L.V., Hafner, I., Fleischhacker, A., Radl, J., Glock, B., Wastian, M., Rößler, M., Lettner, G., Popper, N., Corcoran, K., 2020. Active consumer participation in smart energy systems. Energy and Buildings 227, 110359. doi:https://doi.org/10.1016/j.enbuild.2020.110359.
Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., Kowald, D., 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine , e70002.
Sharma, K., Kaur, A., Gujral, S., 2014. Brain tumor detection based on machine learning algorithms. International Journal of Computer Applications .
Shekhorkina, S., Babenko, M., Spyrydonenkov, V., Shevchenko, T., 2024. Scalability and replicability analysis for an intelligent building management system, in: International Scientific Conference EcoComfort and Current Issues of Civil Engineering, Springer. pp. 483–495.
Snoonian, D., 2003. Smart buildings. IEEE spectrum , 18–23.
Tripathi, M.K., Maktedar, D.D., 2016. Recent machine learning based approaches for disease detection and classification of agricultural products, in: 2016 international conference on computing communication control and automation (ICCUBEA), IEEE. pp. 1–6.

Varoquaux, G., Colliot, O., 2023. Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders , 601–630.

Verticchio, E., Martinelli, L., Gigliarelli, E., Calcerano, F., 2024. Current practices and open issues on the whole-building dynamic simulation of historical buildings: A review of the literature case studies. Building and Environment , 111621.

Xiao, F., Zhao, Y., Wen, J., Wang, S., 2014. Bayesian network based fdd strategy for variable air volume terminals. Automation in Construction , 106–118.

Yang, H., Zhang, T., Li, H., Woradechjumroen, D., Liu, X., 2014. Hvac equipment, unitary: Fault detection and diagnosis. Encyclopedia of Energy Engineering and Technology , 854–864doi:10.1081/E-EEE2-120051345.

Zhang, J., Zhang, C., Lu, J., Zhao, Y., 2025. Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning. Applied Energy , 124378.