Applying AI to HVAC systems: A critical analysis of data-related challenges

Dalia Ali, Lithuanian Energy Institute
Violeta Motuzienė, Vilnius Gediminas Technical University

Breslaujos St. 3,
Kaunas, Lithuania.
dalia.ali@lei.lt

The lack of accessible, reliable, and audited data is a pressing issue, hindering research progress.

Nearly 49% of the studies were excluded from this analysis due to unavailable data, highlighting a major barrier to AI-driven HVAC research replication and validation.

In the analyzed studies, 33% used publicly available datasets, while 67% offered data only upon request, often with access delays and restrictions. This highlights the need for more open and easily accessible datasets to support faster and more collaborative AI research in HVAC systems.

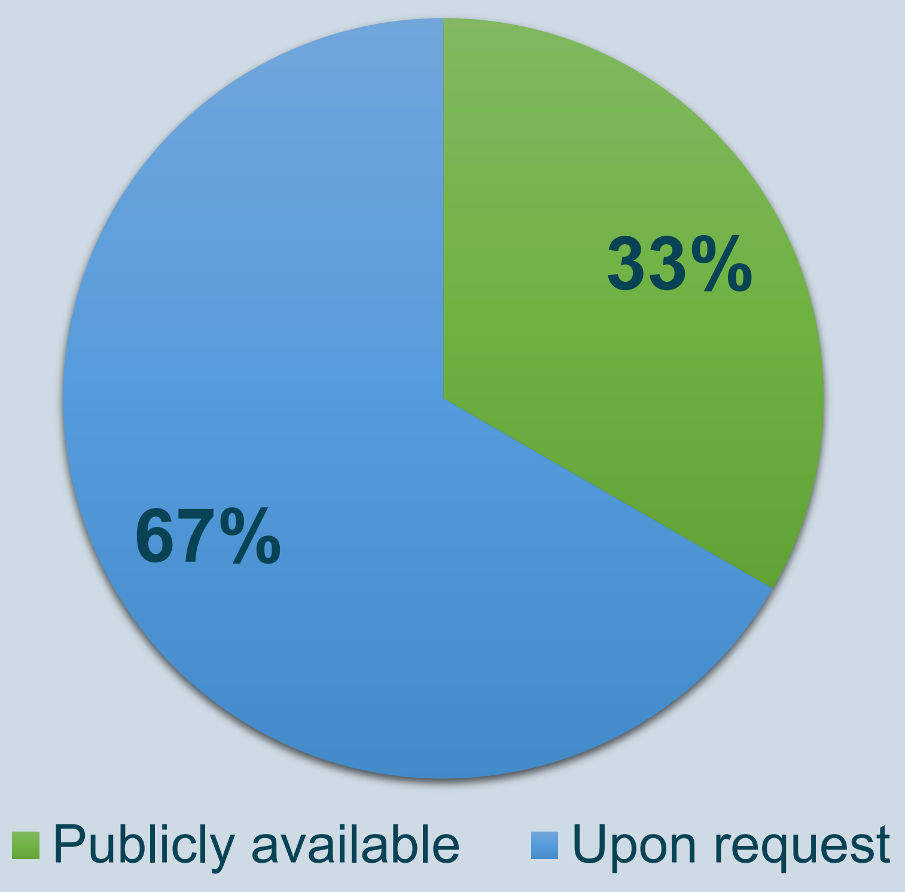


Figure 1. Data accessibility

Open data source platforms like Kaggle offer HVAC datasets, but many lack reliability or completeness, which affects AI model performance. Accurate, audited data remains limited. To support robust AI development, we recommend the following trusted datasets in Table 1.

Table 1. Recommendation of audited data sources

Data Source	Description
ASHRAE	The methods used to collect, process, and analyse ASHRAE datasets are reviewed to ensure scientific credibility and accuracy.
U.S. Department of Energy’s Office of Scientific and Tech. Information	Datasets provided by the US Department of Energy's Office of Scientific and Technical Information (OSTI) are generally reviewed and validated.
HARMONAC	The project involves extensive data collection from different buildings across Europe to assess HVAC performance and identify potential energy savings.

Introduction

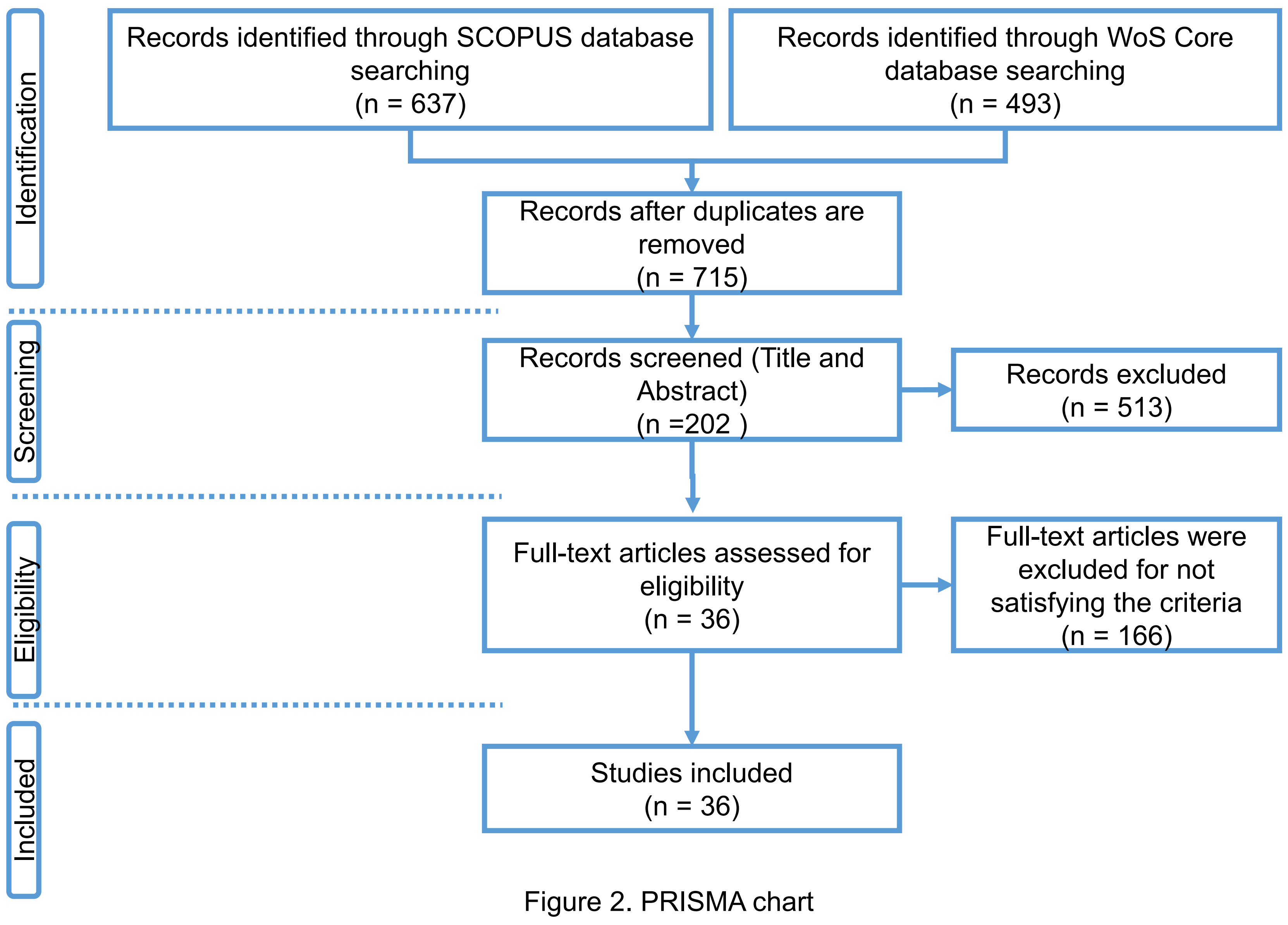
Heating, Ventilation, and Air Conditioning (HVAC) systems consume over 30% of building energy worldwide, making them a critical target for optimization. Artificial Intelligence (AI), especially deep learning and hybrid models, offers powerful tools for improving energy efficiency through smart control, fault detection, and predictive maintenance. However, the effectiveness of AI relies heavily on the quality, availability, and management of data. This study analyzes recent AI-based HVAC research from a data perspective, highlighting key challenges such as limited real-world datasets, data preprocessing issues, and a lack of standardized practices. It recommends improving data reliability and accessibility to enhance model performance and real-world applicability.

Methodology

The literature selection criteria were as follows:

- Recent (<5 years).
- Includes DL or a hybrid model.
- Data is available publicly or upon request.
- Not a review paper.
- Doesn’t involve renewable energy systems.

The figure below outlines the systematic approach taken to identify, screen, and select the studies included in this analysis



Records identified through SCOPUS database searching
(n = 637)

Records identified through WoS Core database searching
(n = 493)

Records after duplicates are removed
(n = 715)

Records screened (Title and Abstract)
(n =202)

Records excluded
(n = 513)

Full-text articles assessed for eligibility
(n = 36)

Full-text articles were excluded for not satisfying the criteria
(n = 166)

Studies included
(n = 36)

Figure 2. PRISMA chart

Results & Conclusion

AI-driven HVAC applications use three main data types: historical, real-time, and simulated data. Historical data is most common due to its availability but relying on it alone limits adaptability. Real-time data offers responsiveness but is complex to implement, while simulated data supports testing when real data is limited. Combining multiple data types enhances model robustness, adaptability, and accuracy in real-world HVAC scenarios.

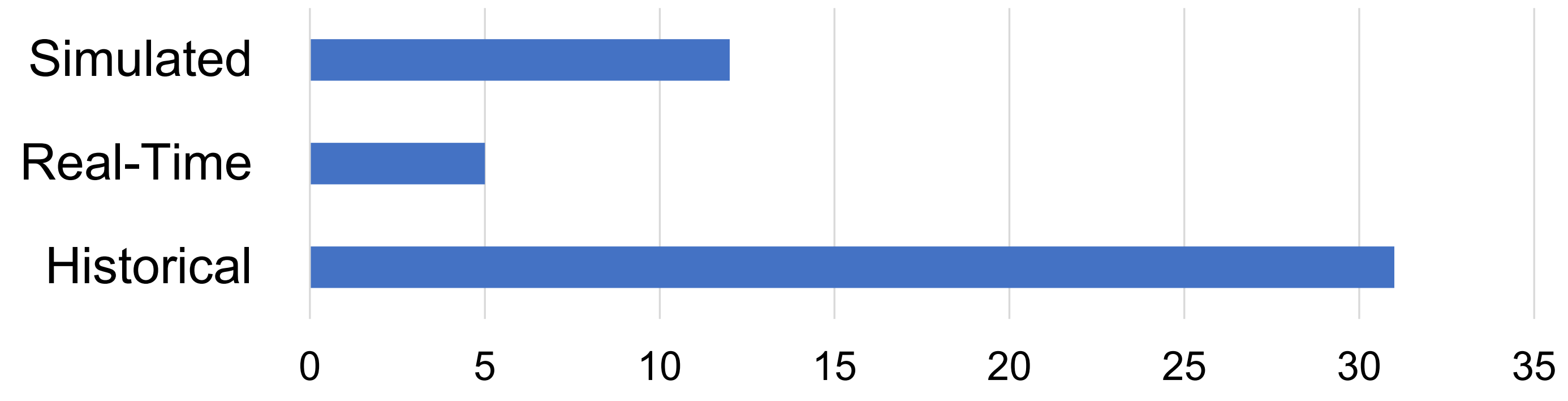
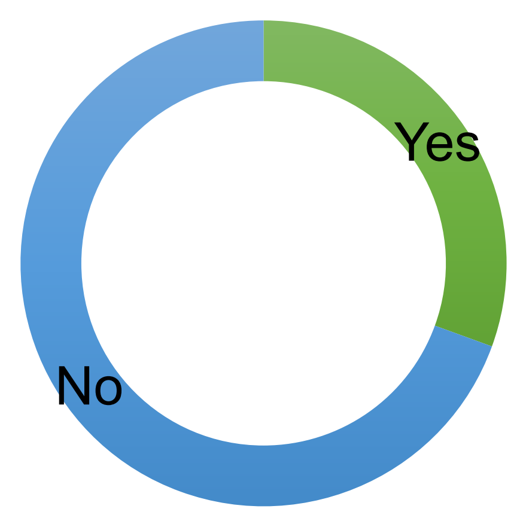


Figure 3. Number of papers by data type

Data quality greatly affects AI model performance, yet 69% of the analyzed studies did not address quality issues, indicating it is not a standard practice in HVAC research. While a few papers report challenges like missing data, noise, and outliers, and apply methods like imputation or cleaning, most overlook this critical aspect, highlighting a gap in current research practices.



Yes – Studies that discussed data quality issues
No – Studies that did not discuss data quality issues

Figure 4. Proportion of studies discussing data quality issues

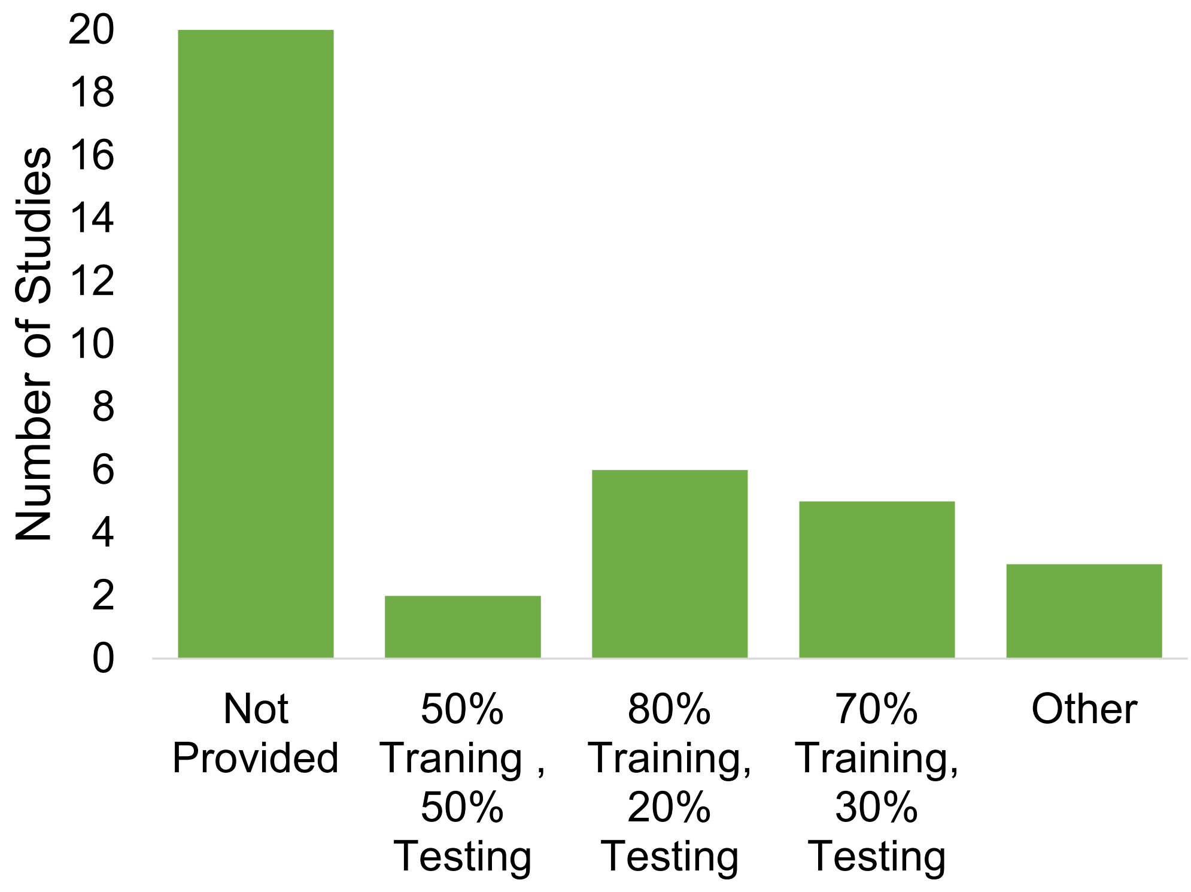


Figure 5. Data split approaches in the studies

Differences in data splitting approaches between studies highlight the need for careful method selection and clear reporting to ensure reliable model evaluation and comparability.