buildings

MDPI

Review

Machine Learning Techniques Focusing on the Energy Performance of Buildings: A Dimensions and Methods Analysis

Maria Anastasiadou [1,*], Vítor Santos [1] and Miguel Sales Dias [2]

[1] NOVA Information Management School, Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal; vsantos@novaims.unl.pt

[2] Department of Information Science and Technology, Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal; miguel.dias@iscte-iul.pt

* Correspondence: manastasiadou@novaims.unl.pt

Abstract: The problem of energy consumption and the importance of improving existing buildings' energy performance are notorious. This work aims to contribute to this improvement by identifying the latest and most appropriate machine learning or statistical techniques, which analyze this problem by looking at large quantities of building energy performance certification data and other data sources. PRISMA, a well-established systematic literature review and meta-analysis method, was used to detect specific factors that influence the energy performance of buildings, resulting in an analysis of 35 papers published between 2016 and April 2021, creating a baseline for further inquiry. Through this systematic literature review and bibliometric analysis, machine learning and statistical approaches primarily based on building energy certification data were identified and analyzed in two groups: (1) automatic evaluation of buildings' energy performance and, (2) prediction of energy-efficient retrofit measures. The main contribution of our study is a conceptual and theoretical framework applicable in the analysis of the energy performance of buildings with intelligent computational methods. With our framework, the reader can understand which approaches are most used and more appropriate for analyzing the energy performance of different types of buildings, discussing the dimensions that are better used in such approaches.

Keywords: energy performance certificate (EPC); machine learning (ML); energy-efficient retrofitting measures (EERM); energy performance of buildings (EPB); energy efficiency (EE)

check for updates

Citation: Anastasiadou, M.; Santos, V.; Dias, M.S. Machine Learning Techniques Focusing on the Energy Performance of Buildings: A Dimensions and Methods Analysis. *Buildings* **2022**, *12*, 28. https://doi.org/10.3390/buildings12010028

Academic Editors: Roberto Alonso González-Lezcano, Francesco Nocera and Rosa Giuseppina Caponetto

Received: 3 December 2021
Accepted: 23 December 2021
Published: 31 December 2021

1. Introduction

Considering that buildings account for 40% of the primary energy consumption (EC) in the European Union [1], reducing the EC of buildings has become a necessity. The European Union, considering the increasing urbanization and climate change trends, defined the objective to reduce EC by 32.5% until 2030, from the baseline year of 2007, as a key priority in the EU's strategy and Green deal [2] to increase EE and decrease the energy performance (EP) of existing buildings [2–4]. This goal is aligned with the United Nations' seventh Sustainable Development Goal (SDG): "Ensure access to affordable, reliable, sustainable and modern energy for all" [5].

Buildings are responsible for the second largest portion of the final EC in the European Union [1,6,7], with households on 26.3% and public buildings on 28.8%, just after the transport sector (with 30.9%). Their refurbishment and energy-efficient retrofitting is a priority for many countries to reduce EC and decrease the EP of existing buildings as part of the EU Green deal [2,8]. In the current state of the art, data science and machine learning are available to analyze, predict and improve energy efficiency (EE) in buildings in meaningful ways. Such computer science approaches can be used to forecast and minimize energy consumption, design energy-efficient buildings, define strategies for mitigating impacts on the environment and climate, and predict and propose useful and cost-effective retrofit measures to increase the EE of buildings to provide a comfortable indoor living environment [9,10]. By measuring, monitoring, and improving the EE in buildings, we can reduce the amount of energy consumed while maintaining or even enhancing the quality of services provided by those buildings, a "double the global rate of improvement in EE"—SDG7.8 [5,11].

This paper proposes a conceptual and theoretical framework applicable in the analysis of literature papers that tackle the problem of the EPB with machine learning or statistical methods. In more detail, this work aims to add to the improvement of the EP of existing buildings, one of the core goals of the EU Green deal [2], by identifying and analyzing the latest and most appropriate machine learning or statistical techniques, as a baseline for future research by building a conceptual and theoretical framework based on a systematic literature review using PRISMA guidelines. Our approach helps the researcher find which methods are most used and more appropriate for analyzing the EP of different types of buildings.

Moreover, our framework addresses the dimensions and factors extracted from available data sources such as building energy certification data, EC data, wheatear and climate data, and others. Our proposal will help the community foster innovation on enhanced buildings' energy performance (EP) and predict energy-efficient retrofit measures (EERM).

In this context, our study adopts a well-established systematic literature review (SLR) method, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA [12]), to identify the most relevant literature contributions to the energy performance of buildings (EPB) and the prediction of EERM, using machine learning (ML) or statistical methods. Furthermore, we used a visualization bibliometric tool, VOSviewer [13], to find the most used terms in the literature related to the EPB with machine learning or statistical methods.

Some literature review papers tackle similar problems, mostly related to EC [14–18]. The main innovation and novelty of the study is how we present and group the data, focusing on the building types and addressing the dimensions and methods for each type. We believe that our study will help the community foster innovation on the enhanced EPB and predict energy-efficient retrofit measures. We present and visualize our results using the bibliometric network software tool VOSviewer. This tool allows creating and visualizing bibliometric networks based on text data and keyword co-occurrence, and authors' co-authorship networks of terms. This allows us to visualize and identify the most important terms and authors co-authorship respective relations for quantitative analysis.

Considering the stated intentions of this paper, we raised the following research questions:

- RQ1: What are the most relevant machine learning or statistical approaches that automatically evaluate buildings' EP using EPC data?
- RQ2: What are the most relevant machine learning or statistical approaches for predicting energy-efficient retrofit measures to improve buildings' EP?

The research questions focus on two objectives (1) automatic clustering—classification of the EPC of a building, and (2) prediction of energy-efficient retrofit measures, using ML and EPC data. Additionally, as mentioned, our approach brings a clear contribution to the EU Green deal and SDG7 of the United Nations [5].

Our paper is organized as follows. Section 2 presents the adopted systematic literature review technique (PRISMA) and our overall methodology. Section 3 describes the application of PRISMA and details the collected data from the survey, whereas in Section 4, we present and analyze such results using the visualization and bibliometric tool. Section 5 discusses our findings, aligned with our research questions, while in Section 6, we present our conclusions.

2. Methodology

The SLR analysis was performed by adopting a well-established systematic literature review and meta-analysis method (PRISMA). In our methodology, we combined this method with data visualization techniques, ending up with 4 main phases: (1) data selection, (2) results and analysis: survey results, categorization and dimensions analysis, visualization and bibliometric analysis, (3) discussion, (4) conclusions [19], as depicted in Figure 1.

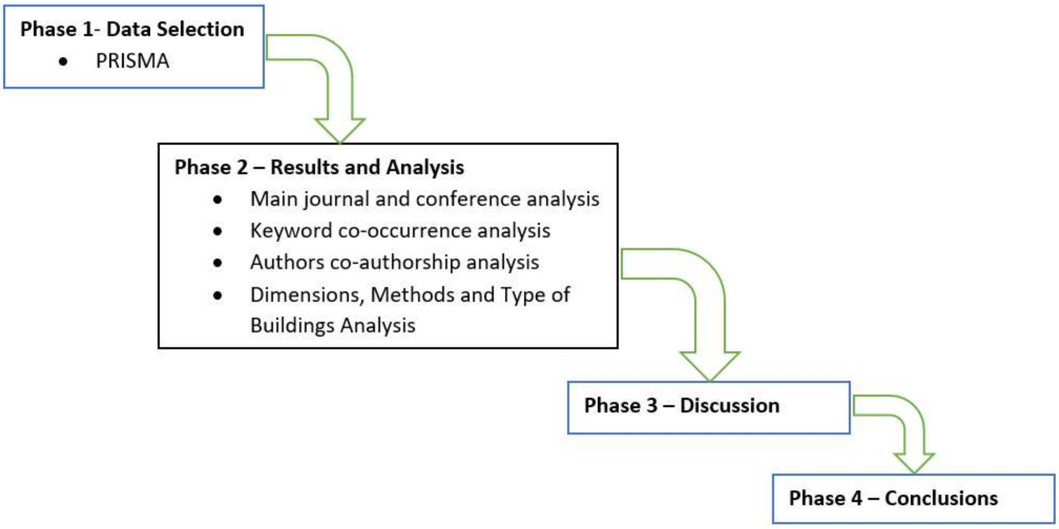
Phase 1- Data Selection
• PRISMA
Phase 2 – Results and Analysis
• Main journal and conference analysis
• Keyword co-occurrence analysis
• Authors co-authorship analysis
• Dimensions, Methods and Type of Buildings Analysis
Phase 3 – Discussion
Phase 4 – Conclusions

Figure 1. Methodology.

Phase 1—Data Collection: Following PRISMA guidelines [12], we conducted an evidence-based systematic review to select the best basis for reporting systematic reviews. Our adoption of PRISMA follows the literature trend of using such a method as a basis for reporting systematic reviews, especially evaluations of interventions [12]. The PRISMA guidelines consist of a flow diagram and a checklist. The flow diagram of conducting a PRISMA survey has four phases: identification, screening, eligibility, and inclusion, as depicted in Figure 2. The checklist proposes a pre-defined structure for a survey with different sections. In addition, there are precise guidelines to be followed and described in more detail in Section 3 [12]. As mentioned, we focused our analysis on ML or statistical approaches using the public build, residential, and office buildings.

Phase 2—Results and Analysis: In this phase, we present the analysis of our PRISMA results. We analyze the main journals and conferences, the keyword co-occurrence, and the authors' co-authorship. We present and visualize our results using the bibliometric network software tool VOSviewer. This tool allows creating and visualizing bibliometric networks based on text data, particularly keyword co-occurrence and authors' co-authorship networks of terms. This analysis illustrates the relationships and connections between the network's elements (nodes), corresponding to the most used terms, allowing the identification of networks characteristics, such as node and cluster centrality. VOSviewer calculates the node links and weight, demonstrating each node's importance in the network. This allows us to visualize and detect the most important terms and authors' co-authorship individual relations for quantitative analysis. The size of nodes presents the degree of centrality: the larger the node, the more times it is reported in the text data. The thickness of edges presents the number of times two linked nodes are reported, showing their significance; by default, the networks are allocated from the largest to the smallest [13]. With this approach, we could summarize and critically analyze the most used dimensions, clustering and classification techniques, EP retrofitting prediction techniques, and the most used building types in each study. This method allowed us to find, accurately and efficiently, the best literature modeling practices and techniques for achieving enhanced EP.

Phase 3—Discussion: In this phase, we discuss the previous phases' findings by following the research questions. We specifically address the identified knowledge gaps and our study limitations.

Phase 4—Conclusion: We sum up and present the conclusion of our study.

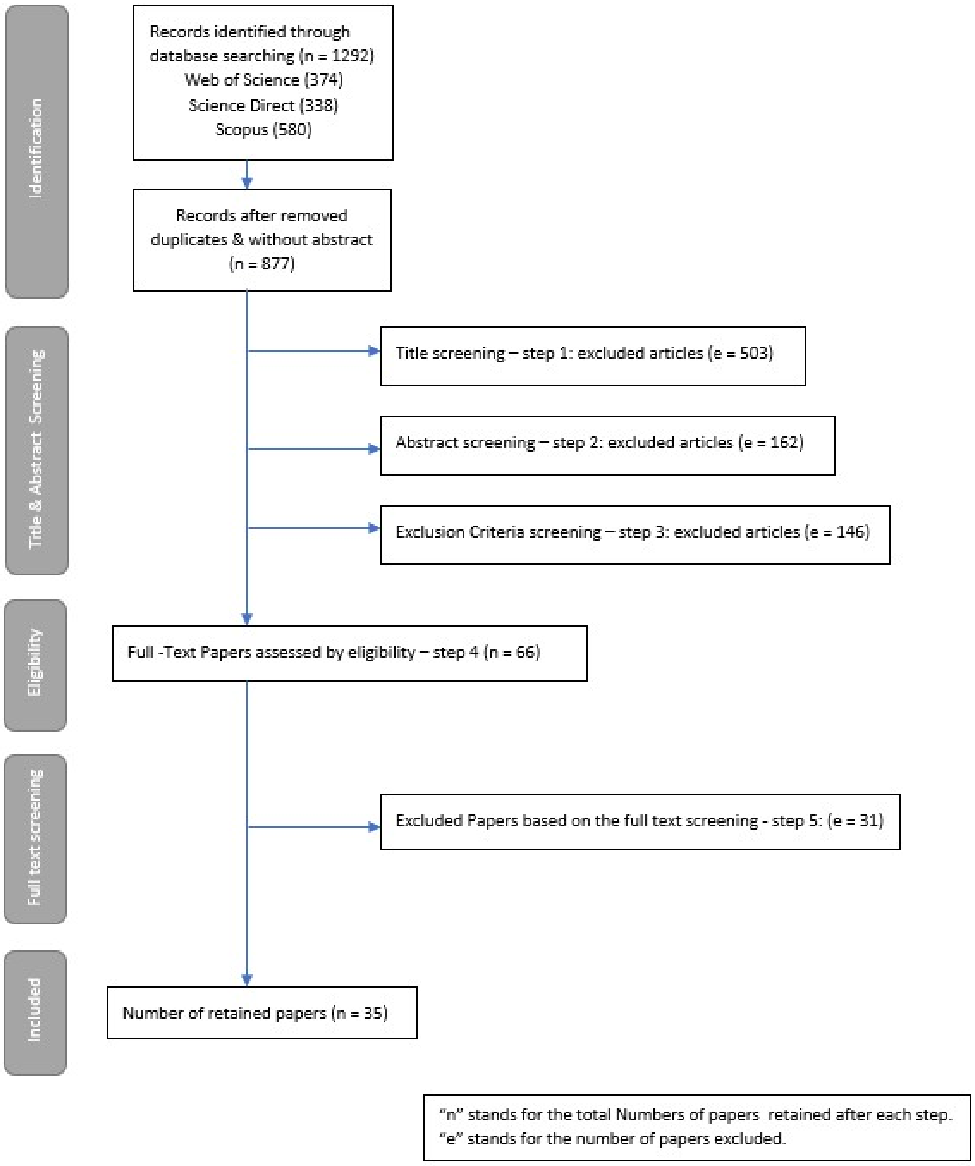
Identification
Title & Abstract Screening
Eligibility
Full text screening
Included
Records identified through database searching (n = 1292) Web of Science (374) Science Direct (338) Scopus (580)
Records after removed duplicates & without abstract (n = 877)
Title screening – step 1: excluded articles (e = 503)
Abstract screening – step 2: excluded articles (e = 162)
Exclusion Criteria screening – step 3: excluded articles (e = 146)
Full -Text Papers assessed by eligibility – step 4 (n = 66)
Excluded Papers based on the full text screening - step 5: (e = 31)
Number of retained papers (n = 35)
"n" stands for the total Numbers of papers retained after each step.
"e" stands for the number of papers excluded.

Figure 2. PRISMA Flowchart.

3. Data Collection

3.1. PRISMA Method

By adopting the PRISMA guidelines, the SLR was performed as follows. First, a search process was conducted to detect publications that have in their titles, abstract or keywords the following Boolean expressions:

> *("energy retrofit*" OR "energy performance" OR "energy analysis" AND ("artificial intelligence" OR "artificial neural networks" OR "machine learning" OR "genetic algorithms" OR "classification" OR "clustering analysis") OR "certificat*" OR "hypercube" OR "k-means")*

The literature search was performed in April 2021 using the following data repositories: Science Direct, Web of Science, and Scopus. Using 'OR,' and 'AND' statements, we include all papers published between the periods 1st January 2016–27th of April 2021. The analyzed topics were integrative, including computer science, mathematics, engineering, environmental, and data science. While all sources were used, the analysis indicated that most of the publications from Science Direct were also in Web of Science and Scopus.

The final set of SLR papers for qualitative and quantitative analysis was organized using the Mendeley references manager open-source tool [20]. This step permitted us to extract metadata, remove duplicates, and obtain precise figures on the relative importance of the author of a particular keyword. The obtained metadata were: authors, publication metadata, references, and citations.

3.2. PRISMA Results

The following PRISMA flow diagram presents the SLR data collection process for our quantitative and qualitative analyses (Figure 2). The initial step in this approach identified published papers through a database search, resulting in 1292 publications (Web of Science 374; Science Direct 338; Scopus 580). The inclusion criteria were original research papers written in English and published in Q1–Q2 peer-reviewed journals (based on scimago rank) and related conferences in the said period. We focused only on papers with studies within the EU, given the applicability of EU directives and regulations and building energy certification, which differs for countries outside the EU. Moreover, even within the EU, there is variation in the methods used to identify and assess EC and building energy certification [21]. Additionally, review, position, and reports papers were excluded.

Subsequently, we removed duplicates (e = 415). Then, we performed title and abstract screening. Step 1 excluded all the papers whose title was not relevant to the scope and objectives of this study (e = 503). Step 2 excluded all the papers without an abstract or whose abstract was not relevant to the scope and objectives of this study (e = 162). Finally, step 3 excluded all the papers according to the outlined inclusion and exclusion criteria (e = 146) as mentioned in the previous paragraph. Next, the full texts of the remaining 66 papers were read, assessed, and fitted on the scope of the research. Thirty-one papers were excluded, given that they did not use ML or statistical techniques. Finally, the remaining 35 papers were considered eligible for further analysis. Thirty-three were published in scientific journals, whereas two were published in conference proceedings.

4. Results and Analysis

4.1. Journals and Conferences Analysis

In the study of a total of 33 literature papers, we analyzed 13 journal papers, including from Applied Energy (9), Energy & Buildings (7), Sustainable Cities & Society (4), Energies (3), Energy (2), Sustainability (1), IEEE Transactions on Automation Science & Engineering (1), Renewable & Sustainable Energy Reviews (1), Measurement (1), Croatian Review of Economic, Business & Social Statistics (1), Journal Electronics (1), Energy Policy (1) and Neural Computing & Applications (1). Table 1 shows the summary of the journals with their information, that most journals are Q1-quartile ranked (9), representing 90%, (2) are Q2-quartile-ranked, and the remaining (1) is not yet classified by the quartile-ranked [22], although the quartile rank can change over time.

The five major research areas found in the analysis were energy, engineering, environmental science, mathematics, and social sciences. The 33 selected papers' publishers originate from five countries, with most of them from the United Kingdom (6) and The Netherlands (2), followed by Switzerland (2), the United States of America (1), Croatia (1), and China (1). The top publishers found are Elsevier BV (5), Elsevier Ltd. (4), Taylor and Francis Ltd. (2), MDPI Multidisciplinary Digital Publishing Institute (1), MDPI AG (1), Institute of Electrical and Electronics Engineers Inc (1), Croatian Statistical Association (1), Science Press (1) and Springer London (1).

Table 1. Journals details.

Journals	No.	Publisher	Country	Field Publisher
Applied Energy	9	Elsevier BV	United Kingdom	Energy, Engineering, Environmental Science
Energy and Buildings	7	Elsevier BV	Netherlands	Engineering
Sustainable Cities & Society	4	Elsevier BV	Netherlands	Energy, Engineering, Social Sciences
Energies	3	MDPI Multidisciplinary Digital Publishing Institute	Switzerland	Energy, Engineering, Mathematics
Energy	2	Elsevier Ltd.	United Kingdom	Energy, Engineering, Environmental Science, Mathematics
Sustainability	1	MDPI AG	Switzerland	Energy, Environmental Science, Social Sciences
IEEE Transactions on Automation Science and Engineering	1	Institute of Electrical & Electronics Engineers Inc.	United States	Engineering
Renewable & Sustainable Energy Reviews	1	Elsevier Ltd.	United Kingdom	Energy
Measurement	1	Taylor & Francis Ltd.	United Kingdom	Mathematics, Social Sciences
Croatian Review of Economic, Business & Social Statistics	1	Croatian Statistical Association	Croatian	Statistics
Journal Electronics	1	Science Press	China	Engineering
Energy Policy	1	Elsevier BV	United Kingdom	Energy, Environmental Science
Neural Computing & Applications	1	Springer London	United Kingdom	Computer Science

The conferences found in this study were IEEE International Conference on Internet of Things and Green Computing & Communications and Cyber, Physical & Social Computing and Smart Data (2017), and IOP Conference Series: Earth & Environmental Science (2019). Table 2 presents that the major research areas of the conference are computer and environmental science in the United Kingdom and Indonesia.

Table 2. Conferences details.

Conference	No.	Publisher Country	Field
IEEE International Conference on Internet of Things and Green Computing and Communications & Cyber, Physical and Social Computing and Smart Data (2017)	1	United Kingdom	Computer Science
IOP Conference Series: Earth and Environmental Science (2019)	1	Indonesia	Environmental Science

4.2. Keyword Co-Occurrence Analysis

Term co-occurrence analysis was conducted utilizing the mentioned text mining tool for network analysis, VOSviewer. The analysis was conducted utilizing a full counting method, encompassing 143 screened terms, with a minimum threshold of two co-occurrences. Of the total 143, only 21 terms were chosen for the analysis (Table 3).

Table 3. Keywords co-occurrence ranked by the link strength.

Keywords	Occurrence	Total Link Strength
Energy Efficiency	7	10
Building Energy Retrofit	4	8
Machine Learning	4	8
Office Buildings	3	8
Building Energy Performance	5	7
Energy Performance Certificate	5	7
Energyplus	3	7
Multi-Objective Optimization	3	7
Genetic Algorithm	2	6
Sensitivity Analysis	2	6
Artificial Neural Networks	2	5
Building Retrofit	3	5
Cluster Analysis	2	5
Energy Simulation	3	5
Energy Retrofitting	2	4
Energy Savings	2	4
Genetic Algorithm (Nsga-Ii)	2	4
Reference Buildings	2	4
Dell'olmo, Jacopo	2	4
Piscitelli, Marco Savino	2	4
Salata, Ferdinando	2	4
Energy Performance Certificates	2	3
Building Sampling	2	2
Fernández Bandera, C	2	2
Ramos Ruiz, G	2	2
Data Exploration	2	1

Most of the analyzed keywords were related to energy efficiency (EE), building energy retrofit, ML, and building energy performance. The top five found keywords were EE (7 occurrences, 10 total link strength), building energy retrofit (4 occurrences, 8 total link strength), ML (4 occurrences, 8 total link strength), office buildings (3 occurrences, 8 total link strength) and building EP (5 occurrences, 7 total link strength).

In keywords of co-occurrence analysis, four clusters (Figure 3) were found with 21 keywords and 50 links. The biggest nodes of each cluster were identified as EE (blue), building EP and EPC (red), office buildings (green), and energy simulation (yellow).

Focusing on the interrelated network of Figure 3 (21 items, 4 clusters, and 50 links), the energy simulation term (yellow cluster) has a connection only with the term energy efficiency (EE) (blue cluster). The building energy performance term (red cluster) has a connection only with the term EE (blue cluster), and the energy performance certificate (EPC) term (red cluster) has a connection with the terms building retrofit (blue cluster) and energy retrofitting (green cluster). The office buildings term (green cluster) relates to all the clusters, namely with the term's sensitivity analysis (yellow cluster), building energy retrofit and artificial neural networks (ANN) (red cluster), EE, and building retrofit (blue cluster). Finally, the EE term (blue cluster) relates to all the clusters too, namely with the term's energy simulation (yellow cluster), ANN, building energy performance and ML (red cluster), office buildings, and clustering analysis (green cluster).

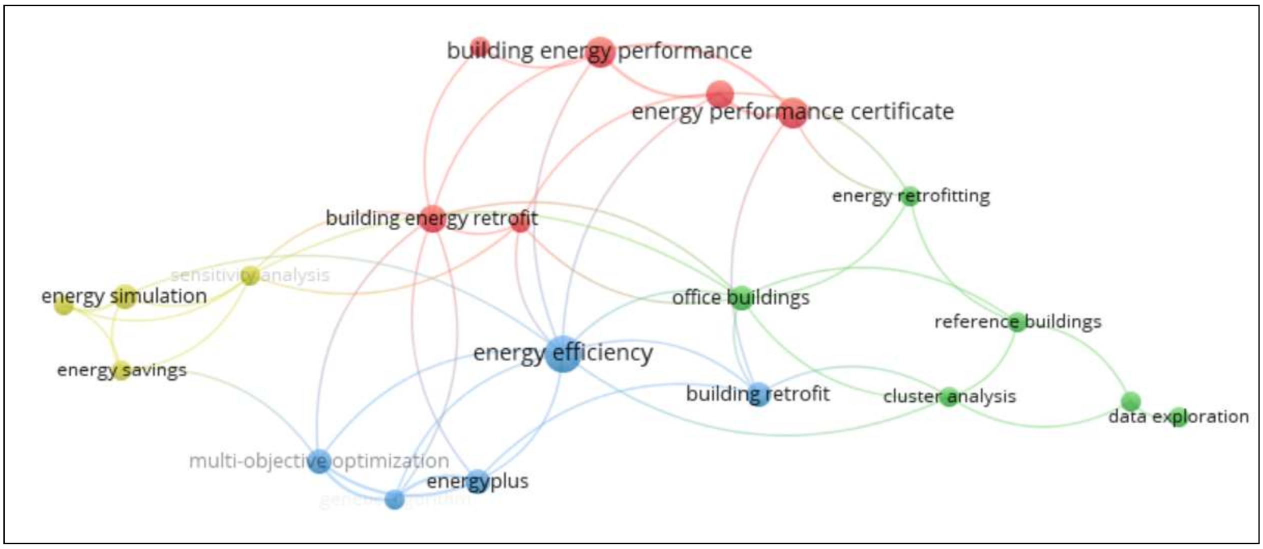

Figure 3. Keyword occurrence network visualization.

An extensive, connected network of keywords and groups of keywords occurs in individual articles, mostly between 2018 and 2020 (Figure 4). The keyword analysis indicated research fields emphasizing ML and EE and found ML techniques, such as clustering analysis, energy simulation, and ANN.

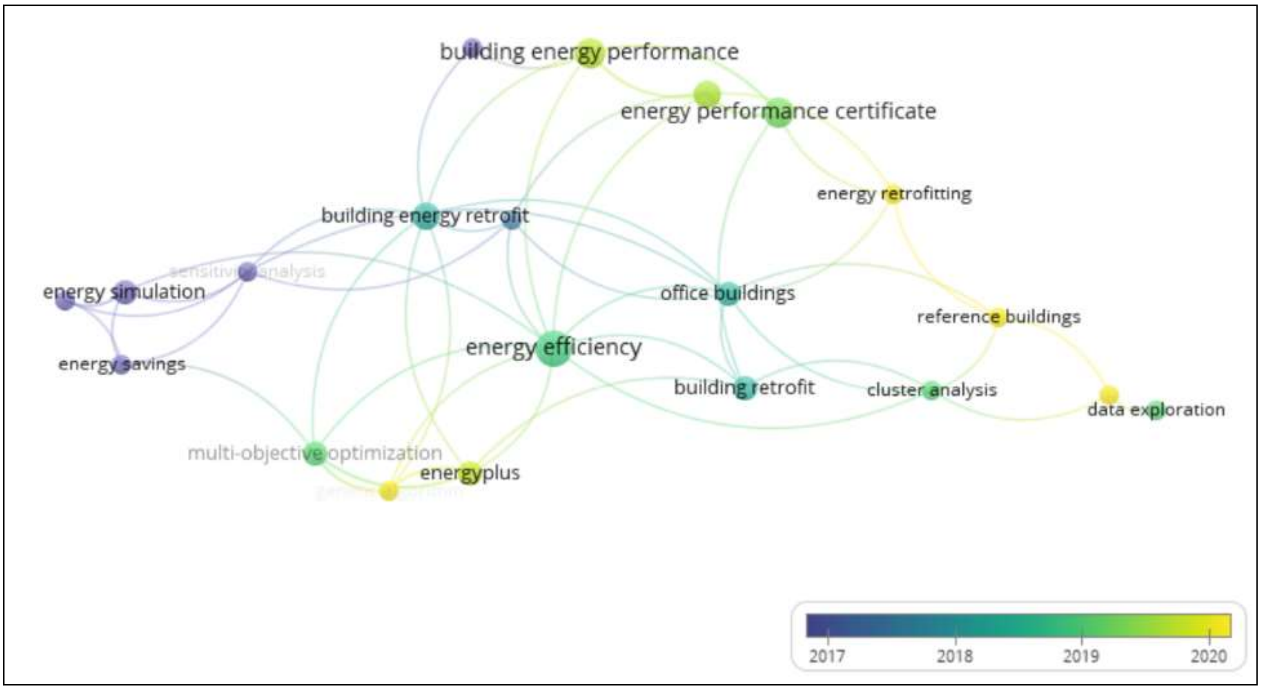

Figure 4. Keyword co-occurrence by year overlay visualization.

4.3. Authors' Co-Authorship Analysis

An authors' occurrence analysis was conducted using the reported text mining tool for network and bibliometric analysis, VOSviewer. The analysis was conducted applying the full count method, choosing 15 maximum number of authors per document and a minimum threshold of 2, resulting in a total of 154 authors meeting this threshold, of which 28 authors were analyzed (Figure 5).

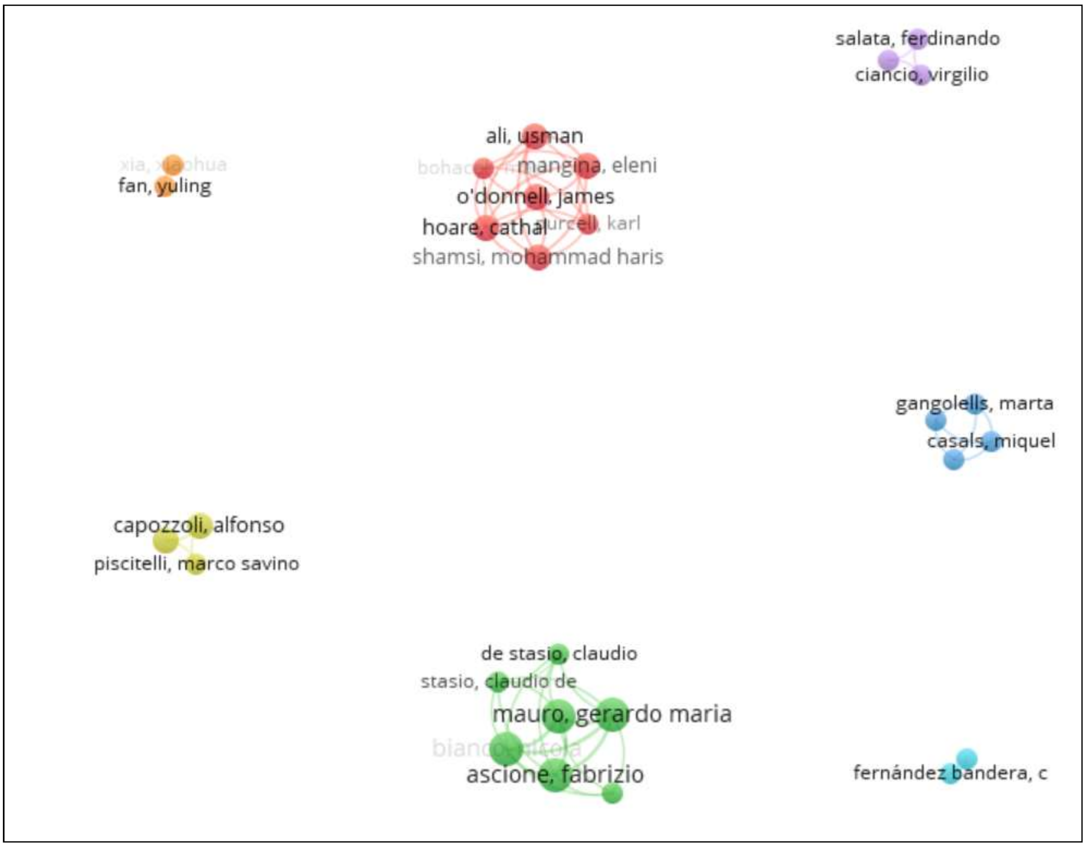

Figure 5. Authors' co-authorship network visualization analysis.

The top 10 found authors were Ascione Fabrizio with a link strength of 21 [23–27], Bianco Nicola with a link strength of 21 [23–27], Mauro Gerardo Maria with a link strength of 21 [23–27], Vanoli Giuseppe Peter with a link strength of 21 [23–27], Ali Usman with a link strength of 16 [28–30], Hoare Cathal with a link strength of 16 [28–30], Mangina Eleni with a link strength of 16 [28–30], O'Donnell James with a link strength of 16 [28–30], Shamsi Mohammad Haris with a link strength of 16 [28–30], and Bohacek Mark with a link strength of 12 [28,29].

In the authors' co-authorship analysis, seven clusters were found with 28 items and 54 links. Cluster 1 (green) relates to the top four author co-authorships ranked by link strength. For De Stasio Claudio with a link strength of 9 [23–25] and De Masi Rosa Francesca with a link strength of 8 [26,27] (Table 4), Cluster 2 (red) has seven items and found Ali Usman and Bohacek Mark. For Hoare Cathal, Mangina Eleni, O'Donnell James, Purcell Karl, Shamsi Mohammad Haris [28–30], Cluster 3 (yellow) has four items and found Casals Miquel, Ferré-Bigorra Jaume, Gangolells Marta, and Macarulla Marcel [31,32]. Cluster 4 (blue) has three items and found Capozzoli Alfonso, Cerquitelli Tania and Piscitelli Marco Savino [33–35], and Cluster 5 (cyan) has three items and found Ciancio Virgilio, Dell'olmo Jacopo and Salata Ferdinando [36,37]. Cluster 6 (purple) has two items: Fernández Bandera C and Ramos Ruiz G [38].

Table 4. Authors' co-authorship ranked by link strength.

Authors	Documents	Total Link Strength
Ascione, Fabrizio	5	21
Bianco, Nicola	5	21
Mauro, Gerardo Maria	5	21
Vanoli, Giuseppe Peter	5	21
Ali, Usman	3	16
Hoare, Cathal	3	16
Mangina, Eleni	3	16
O'Donnell, James	3	16
Shamsi, Mohammad Haris	3	16
Bohacek, Mark	2	12
Purcell, Karl	2	12
De Stasio, Claudio	3	9
De Masi, Rosa Francesca	2	8
Casals, Miquel	2	6
Ferré-Bigorra, Jaume	2	6
Gangolells, Marta	2	6
Macarulla, Marcel	2	6
Capozzoli, Alfonso	3	5
Cerquitelli, Tania	3	5
Ciancio, Virgilio	2	4
Dell'olmo, Jacopo	2	4
Piscitelli, Marco Savino	2	4
Salata, Ferdinando	2	4
Fernández Bandera, C	2	2
Ramos Ruiz, G	2	2

Clusters 2, 3, and 5 relate to authors' published articles in 2020–2021. Cluster 4 relates to authors with publications in 2019, and cluster 7 corresponds to authors with publications in 2018; for the remaining authors, articles were published in 2017. Figure 6 indicates that the top 10 author co-authorships were published in 2017, demonstrating that the academic community had a strong connection in 2017. Finally, the most relevant papers were published from 2017 to 2020, demonstrating that the academic community has increased.

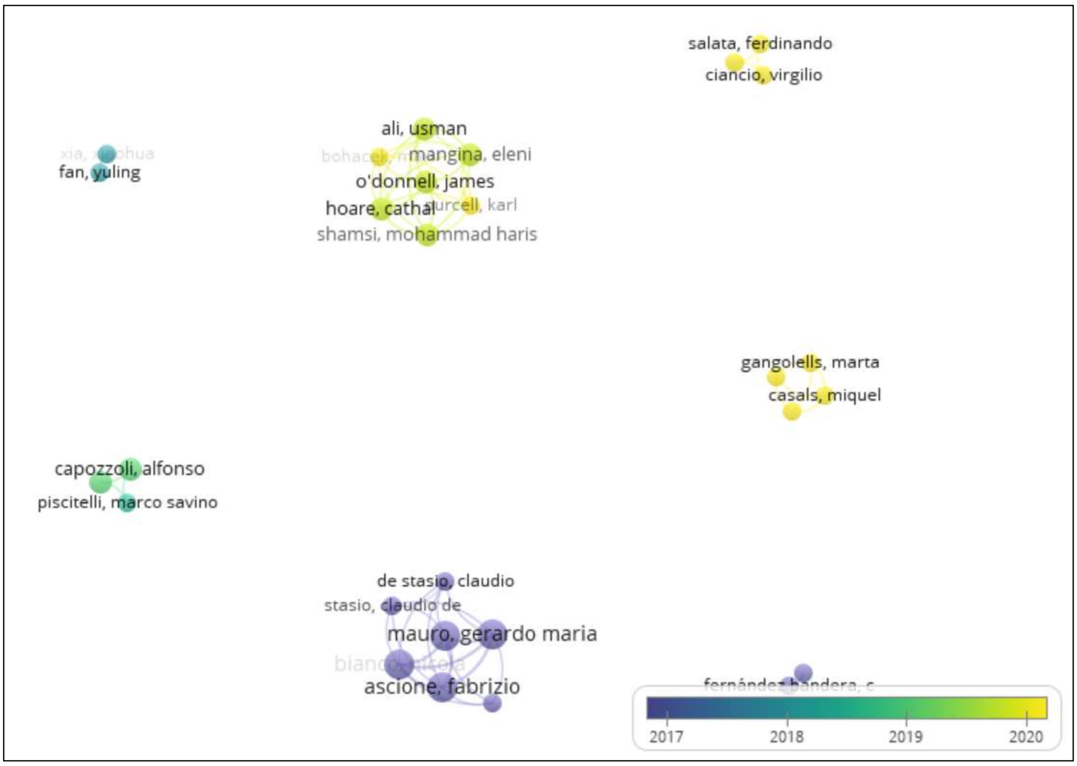

Figure 6. Authors' co-authorship visualization by year.

4.4. Most Cited Publications

Analysis of the most-cited publications helped us to detect the important research topics in the literature. The most cited and chosen publications were searched using Science Direct, Scopus and Web of Science datasets. The study detected publications that have been cited between 84 times and 0 times. Table 5 shows this process's resulting conceptual and theoretical framework with each paper's dimensions, intelligent computing methods, and type of buildings.

The top five found publications are from the following authors: Ascione, Bianco, Stasio et al. [23] with 84 citations (the most cited), followed by Ramos Ruiz et al. [38] with 44 citations, Ascione, Bianco, De Masi et al. [27] with 44 citations, Niemelä, Kosonen, and Jokisalo [39] with 39 citations and Beccali et al. [40] with 37 citations. These results (Table 5) are coherent with previous analyses described above. These papers are the most cited and present the central concepts in the field.

The top five cited papers present in Table 5 were published in Q1-ranked journals and mostly in Energy and Buildings and Applied Energy journals. Furthermore, and coherent to the analysis, the most cited article is also emphasized in the authors' co-authorship analysis (Section 4.3). Cluster 1 (green) in Figure 5 groups the most cited author co-authorship Ascione, Bianco, Stasio et al. [23] and cluster 6 (purple) groups most of the author co-authorships of the second most-cited article Ramos Ruiz et al. [38]. In keyword co-occurrence analysis (Section 4.3), the term ANN was outstanding and is one of the techniques used by the most cited publication of Beccali et al. [40].

Table 5. Publications ranked by the number of citations.

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
1	[16]	Applied Energy	-Thermo-physical characteristics -Building envelope -HVAC systems -Weather -Energy use	-Simulation -Latin hypercube sampling -Pareto—sensitive Analysis -Genetic Algorithm	Hospital	84
2	[31]	Applied Energy	-Climatic location -Geometry -Construction elements -Building properties -Internal temperature measures	-Genetic Algorithm NSGA-II -Simulation -Parametric analysis -Sensitivity analysis -Uncertainty analysis: fi, CV(RMSE)	University	44
3	[20]	Energy & Buildings	-Building envelope -Building operation -HVAC systems -Financial attributes	-Genetic algorithms -Transient energy simulations	University	44
4	[32]	Energy & Buildings	-Building Envelope -HVAC systems -Internal heat gains -Weather -Cost of different renovation measures	-Simulation-based Optimisation methods -Pareto-Archive NSGA-II Genetic algorithm	Residential	39
5	[33]	Energy	-Thermophysical parameters -HVAC plants -Typology -Building characteristics -Climate -Geometry -Energy consumption	-Artificial neural networks (ANN)	School	37
6	[34]	Energy	-Climatic location -Building materials -Financial attributes	-Life-Cycle Cost method -Monte Carlo simulation -Discount rate	Residential	35
7	[35]	Energy & Buildings	-Design variables -Climate (Thermal zone) -Cooling and heating	-Pareto front -Simulation -Non-dominated Sorting Genetic Algorithm-II (NSGA-II)	Residential	30

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
8	[36]	Applied Energy	-Geometry -Weather -Construction materials	-Energy Simulation -Residual network model	Residential	27
9	[30]	Applied Energy	-Climate -Building location -Energy sources (gas and electricity) -Building characteristics -Installation systems -Photovoltaic -Thermal solar panels -Building geometry	-Simulations -Active Archive Non-Dominated Sorting Genetic Algorithm (aNSGA-II type). -Pareto frontier	Residential	22
10	[37]	Applied Energy	-Weather -Building Envelope -HVAC systems -Energy use	-Latin-hypercube sampling -Joint mutual information maximization -Energy conservation measure	Residential and offices	21
11	[38]	Measurement	-Building envelope -Orientation -Heating load -Cooling load	-Estimation Maximization algorithm -Adaptive Neuro-Fuzzy Inference System method -Principal Component Analysis	Residential	14
12	[22]	Applied Energy	-Building geometry -Energy performance index -Building shape -Dwelling type -Building envelope -Number of floors; walls, and windows -Envelope U-values -Construction assemblies -HVAC systems	-Crude statistical analysis -Visual analysis of statistical representation (Box plots) -Local Outlier Factor (LOF) algorithms -Deep Learning -Rule Induction -Neural Network -Naïve Bayes -Decision Tree -Random Forest -Gradient Boosted Trees -Learning Vector Quantization (LVQ) -9 k-Nearest Neighbors (kNN).	Residential	14

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
13	[29]	Energy & Buildings	-Geometry -Envelope -Useful floor area (m^2) -Building shape -Climate zone Window -Glazing type -Wall insulation -Heating system Heating -Energy source -Cooling system Cooling	-Sorting genetic algorithm (aNSGA-II) -Optimal solution in the R 4 space -Pareto frontier -Simulation	Residential	14
14	[39]	Sustainable Cities & Society	-Building Envelope -Heating systems -HVAC systems -Electricity consumption	-REVIT -Simulation -Genetic algorithm	Residential	13
15	[18]	Sustainability	-Geometry -Building envelope -Building operation -HVAC systems -Climate	-Latin hypercube sampling technique -Simulation-based large-scale uncertainty/sensitivity Analysis of Building EP	Residential, Offices and Schools	12
16	[40]	Energies	-Weather -Building age -Construction year -Building envelope -Heating/ Cooling systems	-Group by building age -Simulation Monte Carlo -Genetic Algorithm NSGA-II	Residential	12
17	[41]	Journal IEEE Transactions on Automation Science & Engineering	-Building intrinsic properties -Occupancy patterns -Environmental conditions	-Artificial neural network -Genetic algorithm -Multi regression analysis -Principal component analysis	Care home	9
18	[28]	Energies	-Geometry -Envelope -Construction year -Average global efficiency for space heating	-Artificial Neural Network -Support Vector Machine -Reduced Error Pruning Tree -Random Forests	Residential	8

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
19	[42]	Applied Energy	-Building envelope -Indoor facilities	-Manual grouping method and 'notch test' data -Generic Algorithm -Mathematics	Offices	7
20	[43]	Renewable & Sustainable Energy Reviews	-Thermophysical properties -Building envelope	-Artificial neural network (ANN) -K-means clustering -Geographic information systems (GIS)	School	7
21	[44]	Applied Energy	-Hot water -Internal heat gain and lighting -Wall to floor and window -Standard Emission Rate -Air infiltration rate -Terminal unit energy -Demand and cooling system efficiency. -Roof wall ratio -Solar radiation on the roof	-Simplified Building Energy Model (SBEM) tool -Sensitivity analysis -Gradient boosted regression trees (GBRT) -Cross-validation -Standard statistical re-sampling method, -Sequential Model-based Algorithm Configuration (SMAC)	Commercial, School	7
22	[45]	Croatian Review of Economic, Business & Social Statistics	-Geospatial -Construction shape -Heating characteristics -Cooling characteristics -Meteorological characteristics -Occupational characteristics -Energetic characteristics	-Artificial neural network (ANN) -K-means clustering -Correlation Analysis -Chi-square tests -Symmetric mean average percentage error -DBSCAN algorithm	Residential, Offices and Schools	6
23	[21]	Applied Energy	-Geometric data: building shape, building type, building fabric, number of floors, window-wall ratios -Non-geometric building: envelope U-values, construction assemblies, Heating Ventilation, Air Conditioning (HVAC) systems properties -EPC data -Building footprint -Building height data	-GIS -Decision Analysis (MCDA) approach -Fuzzy string algorithms -Jaro -Jaro–Winkler -Levenshtein -JaccnaïveNaive Bayes Generalized Linear Model -Logistic Regression -Deep Learning -Decision Trees -Random Forest -Gradient Boosted Trees -Support Vector Machine	Residential	6

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
24	[46]	Sustainable Cities & Society	-Geometrical -Thermophysical features	-Wrapper Feature Selection -Random Forests -K-means Clustering	Schools	5
25	[25]	Energy & Buildings	-Building norm -Window glazing -Climate zone -Cooling system -Useful floor area -Shape factor -Domestic hot water energy source -Heating energy source -Existence of thermal insulation in building envelopes	-K-means Clustering -Correlation analysis -Stepwise regression analysis -Root-mean-square standard deviation -Elbow method	Offices industrial—residential	5
26	[47]	Sustainable Cities & Society	-Climate zone -Building layout -Seasonal efficiency Heat delivery efficiency -Average water inlet temperature Hot water supply temperature -Mechanical ventilation Infiltration -Maximum power consumption -Luminous energy conversion efficiency Schedule -Occupants Lighting	-Dynamic simulation tool -IES Virtual Environment (VE) -Combination packages -Energy Limiting Difference (ELD) assessment factor	Residential	4
27	[48]	Energy & Buildings	-Construction age -Building size -Heating and hot water systems -Heat loss through the building fabric -Climatic location -Operation & occupancy pattern -Heating demand	-Statistical approach -Synthetical Average Building (SyAv) approach identifies	Residential	4

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
28	[23]	Energy & Buildings	-Geometric data -Envelope U-values -HVAC systems -Construction year -Climate zone	-Local Outlier Factor algorithm -K-means clustering -Weighting coefficients -Building national statistics -Building EP Simulation -Geographical Information System (GIS) visualization maps	Residential	2
29	[26]	IEEE International Conference on Internet of Things and Green Computing & Communications & Cyber, Physical & Social Computing & Smart Data (2017)	-Buildings Characteristics -Efficiency of the subsystems for space heating -System efficiency -EP (Normalized primary energy demand for space heating [kWh/m^2], etc.)	-Pearson correlation analysis -Principal component analysis -K-means clustering -Classification and Regression Tree algorithm -Silhouette based indices -Singular value decomposition -Statistics -Boxplot distributions -Generalized association rule	Residential	2
30	[49]	Energy Policy	-Dwelling type -Year of construction -Dwelling size -Occupancy status -Energy class -Surface coefficient of heat exchange -Real energy consumption -Systematic energy source -Heating system type -Region Climatic zone -Urban size -Renovation changes	-Statistical approach -Cost–benefit analysis. -Monte Carlo simulation -Sensitivity analysis -Hierarchical Classification (Ward's criterion)	Residential	2
31	[50]	IOP Conference Series: Earth & Environmental Science	-HVAC systems -Envelope U-values	-ENERFUND tool -Geographical Information System (GIS) visualization	Commercial and Residential	1

Table 5. *Cont.*

N	Ref.	Publication	Dimension Category	Methods	Building Type	No. of Citations
32	[27]	Electronics	-Aspect ratio -Surface area -Floor area -Average u-value of the vertical opaque envelope -Average u-value of the windows -Heating system global efficiency -Construction year	-Density-based spatial clustering of application with noise algorithm (dbscan) -Pearson correlation -Max–min binormalization -Elbow method -K-means -Spatial constrained k-nn -Geospatial maps	Residential	1
33	[51]	Energies	-Building type -Number of stories -Construction year -Heated space per story -Area code -Number of stairwells per Apartment	-Google Street View -ANN -Image recognition -Stepwise regression -Logistic regression (LR) -Support vector machines (SVM)	Residential	1
34	[24]	Sustainable Cities & Society	-Useful floor area (m^2) -Building shape -Climate zone Window -Glazing type -Wall insulation -Heating system Heating -Energy source -Cooling system Cooling	-Statistical approach -Life-cycle energy impact: Calculate the global energy savings -Life-cycle economic impact -Calculate life environmental impact	Offices	1
35	[52]	Neural Computing & Applications	-Useful surface (m^2) -Thermal power (kW) -CO^2 emissions -Primary energy consumption -Opaque enclosures -Holes and skylights	-Statistical approach -Bayesian Gaussian process regression (GPR) -Genetic algorithms (GAs) -Limited-memory Broyden–Fletcher–Goldfarb– Shanno (L-BFGS) optimizers	Residential	0

Likewise, several ML and statistical methods were used for energy applications on SLR papers. The 10 top most-cited papers used a combination of methods, namely simulation techniques, Pareto front, genetic algorithm NSGA-II, and ANN (Table 5). As input in those methods, these top 10 papers used the following dimensions extracted from the data: climate and weather, building thermo-physical characteristics, building envelope, building geometry, HVAC systems, EC, and building typology. In the case studies, most of them used residential buildings (6), offices (1), universities (2), schools (1), and hospitals (1), refs. [23,27,37–44]. The remaining SLR papers used similar dimensions: building geometry, building envelope, other building properties, climate and weather, HVAC systems, and energy consumption (EC) [16,18,20,21,28–48]. A total of 19 out of 35 papers used a residential building as the case study [28–30,33–37,39,41–43,45–47,53–57]. Some of them combine different types of buildings. Five papers combined residential and commercial buildings—offices and schools [25,32,44,52,58], two papers addressed offices [31,49] and six analyzed schools and universities [27,38,40,50,51,59]. Only one addressed a hospital [23] and one a care home [48].

Furthermore, only the most recent papers utilized building EPC data for their analysis [28,29,31,34,50–54,56,57,60,61]. This aspect is surprising since the first directive on building energy performance, "the Energy Performance of Building Directive (EPBD)," was introduced by the European Parliament in 2002. Additionally, improvements to the EPBD were performed in 2010 [60,61]. The remaining papers use energy building audits analysis and reference buildings for their research.

The most-used techniques for predicting EP and retrofitting were energy performance simulation techniques, statistical-based approaches, genetic algorithms, and ANN. Few studies use only ML methods, namely (13) studies [28,29,31,33–35,40,45,48,50,52,56,59]. The most common clustering and classification techniques were K-means (7), statistical methods (6), Latin hypercube sampling (2), other manual groupings (2), decision tree (2), and probability density function (1) [23,25,28–34,47,50–52,54,55,57,59].

4.5. Type of Buildings, Dimensions, and Methods Analysis

A conceptual and theoretical framework was built to evaluate this survey's building types, dimensions, and computational intelligence methods in more detail; see Tables 6 and 7. This framework seeks to understand the most-used ML and statistical approach according to each SLP study's dimensions and building types resulting from the previous analysis (Table 5). It focuses on research inputs, goals, and outcomes to create the basis for our research evaluation criteria.

Table 6. Analysis of the used Dimensions by Type of Buildings.

No.	Building Type	Dimension Category	Reference
1	Hospital	-Thermo-physical characteristics -Building envelope -HVAC systems -Weather -Energy use	[16]
3	University/School	-Building envelope -Building operation -HVAC systems -Financial attributes -Thermophysical parameters -Typology -Climate -Geometry -Energy consumption -Hot water -Internal heat gain and lighting -Standard Emission Rate -Air infiltration rate -Terminal unit energy -Demand and cooling system efficiency. -Roof wall ratio -Solar radiation on the roof	[20,31,33,43,44,46]

Table 6. *Cont.*

No.	Building Type	Dimension Category	Reference
4	Residential	-Building Envelope -Geometry -HVAC systems -Internal heat gains -Weather -Building materials -Financial attributes -Cost data of different renovation measures -Building–location–orientation -Photovoltaic -Energy performance index -Envelope U-values -Construction assemblies -Thermal solar panels -Electricity consumption -Building age -Construction year -Average global efficiency for space heating -EPC -Construction age -EP (Normalized primary energy demand for space heating [kWh/m^2], etc.) -Occupancy status -Energy class -Renovation changes -CO_2 emissions -Primary energy consumption	[21–23,26–30,32,34–36,38–40,47–49,51,52]
4	Residential and offices	-Weather -Building Envelope -HVAC systems -Energy use -Envelope U-values	[37,50]
5	Residential, Offices and Schools	-Geometry -Building envelope -HVAC systems -Climate -Geospatial -Construction shape -Occupational characteristics -Energetic characteristics -Building Insulation	[18,25,45]
6	Care homes	-Building intrinsic properties -Occupancy patterns -Environmental conditions	[41]
7	Offices	-Building envelope -Indoor facilities -Useful floor area (m2) -Building shape -Climate -Glazing type -Wall insulation -Heating system -Energy source -Cooling system	[24,42]

Table 4 presents our findings on dimensions by building types to implicate new knowledge, which helps energy experts to learn and use the most critical dimensions for particular building types in their modeling and research work.

The SLR analysis suggests that the dimensions extracted from the data sources, can be grouped in the following way:

- Climate: location, weather, building orientation.
- Building geometry: building shape, building type, building fabric, number of floors, window-wall ratios.
- Non-geometric building data: envelope U-values, construction assemblies, heating ventilation, air conditioning (HVAC) systems properties, building age.
- Energy consumption: electricity consumption, energy use, average global efficiency for space heating, HVAC systems, internal heat gain, and lighting.

- Energy performance: standard emission rate, CO_2 emissions, terminal unit energy, energy performance index, the efficiency of the subsystems for space heating.
- Financial attributes: cost data of different renovation measures
- Occupational characteristics.

Table 7 presents our inference which may help data scientists understand the right method to employ for further research.

Table 7. Methods by Dimensional Analysis.

No.	Computational Intelligence Method	Dimension Category	Reference
1	Simulation	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance	[16,18,20,23,29–32,34,37,39,40,47,49]
2	Genetic Algorithm	-Climate -Geometric building -Non-geometric building data -Energy consumption -Energy performance -Financial attributes	[16,20,29–32,35,39–42,52]
3	Sensitivity analysis	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Financial attributes	[31,44,49]
4	Artificial neural networks (ANN)	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Occupational characteristics	[28,33,41,43,45,51]
5	K-means clustering	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Occupational characteristics	[23,25–27,43,45,46]
6	Geographic information systems (GIS)	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance	[21,23,27,43,50,51]
7	DBSCAN algorithm	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Occupational characteristics	[27,45]
8	Correlation analysis	-Climate -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Occupational characteristics	[23,25–27,45]

Table 7. *Cont.*

No.	Computational Intelligence Method	Dimension Category	Reference
9	Statistical approach	-Climate data -Building geometry -Non-geometric building data -Energy consumption -Energy performance	[22–26,29,30,35,41,42,44,48,49,51,52]
10	Cost-Benefit analysis	-Climate data -Building geometry -Non-geometric building data -Energy consumption -Energy performance -Financial attributes	[24,34,49]
11	Principal Component Analysis	-Climate -Building geometry -Non-geometric building data -Energy consumption	[38,41]

The above analysis allows us to use the most common dimension categories of building to find an adequate method to evaluate the energy performance according to the building type we are interested in. As the results demonstrated, most studies have common dimensions no matter the building type and methods.

5. Discussion

Our research aimed to highlight and detect the literature on machine learning (ML) and statistical techniques that tackle the EPB and create a systematic, organized view of those literature studies.

Following, we discuss how our study answers the posed research questions, namely:

- RQ1: What are the most relevant machine learning or statistical approaches that automatically evaluate buildings' energy performance using EPC data?
- RQ2: What are the most relevant ML or statistical approaches for predicting energy-efficient retrofit measures to improve buildings' energy performance?

5.1. Research Questions Discussion

Our analysis indicates that the two problems discussed by the proposed machine learning or statistical approaches are clustering (classification) and prediction in the energy performance of buildings.

Regarding the first question (RQ1), 13 studies used the EPC dataset [30,32–35,52,58] as explained in Sections 4.4 and 4.5. This kind of data is multi-dimensional, given that each energy certificate has many attributes. The exploitation of a given data mining algorithm on such data (such as cluster analysis) is challenging due to the high variability and dimensionality of the data [33]. As for data classification and clustering techniques, most studies applied the K-means clustering algorithm to characterize the cluster sets with given energy performance, as explained in Sections 4.4 and 4.5. Some studies used a density-based spatial clustering of application with noise algorithm (DBSCAN) to handle outliers and correlation analysis to identify the best input demission for their clustering analysis [32–34,52]. A few studies referring to RQ1 used GIS and geospatial maps to visualize their clustering results [30,58]. Finally, (5) papers of similar studies answered RQ1, namely [30,33–35,58].

Regarding RQ2, most approaches to predicting energy-efficient retrofit measures used simulation tools such as EnergyPlus [62] or TRNSYS [63] to model the energy consumption (EC) of a 3D model of the building. They understudy and then use GA to perform multi-objective optimization, obtaining a good solution for the different criteria defined as important in their studies [46]. The strategy of using precomputed 3D models requires a large database of models and the accuracy depends on how close those models match real-world buildings. Although the most common algorithm for multi-objective optimization is the non-dominated sorting GA II (NSGA-II), it is possible to improve the algorithm by customizing it for energy retrofitting scenarios [64]. NSGA-II is a GA and customizing it for the specific field of energy retrofit would yield more efficient computations. Additionally, the more recent NSGA-III is not used by researchers [65]. The improved version will be more efficient computationally when finding optimal solutions. The simulation's quality depends on having a good model representation of the building and using other environmental factors such as weather data and orientation of the building/solar exposition [44].

The environmental characteristics that impact the building EP are also important criteria to determine what retrofitting measures are cost-effective. It is also essential to describe the building materials in terms of their heat loss/gain rating by the thicknesses (U-value) of features, namely roof, wall, floor, ceiling, and window, as well as identify the type of heating and cooling systems, renewable energy systems being used, occupation density, and others that might affect the building's energy consumption. It can be considered that the more extreme the weather conditions are in the region of the building, the more critical it is to include it in the modeling of EP [44].

Moreover, referring to the RQ2, several authors used GA (7) [23,27,38,39,42,48] to predict cost-optimal energy retrofit solutions. Some approaches used artificial neural networks (ANN) [35,40,48,50,52,56]. Most papers in this category are case studies using a single building or a representative building sample to collect the necessary data to serve their experiments. No study referring to RQ2 used GIS and geospatial maps to visualize. Finally, (15) papers of similar studies answered the RQ2, namely [23,25,27,29,36–39,41,42,44,46,48,55].

Some studies (8) answer both research questions; two such approaches are an excellent example of using K-means clustering and ANN with public EPC databases to predict EERM [50,52]. Other approaches focusing only on predicting energy consumption (EC) show that it is possible to use a data-driven urban energy simulation to predict the hourly, daily, and monthly energy consumption. In addition, models are used as a baseline for EC and then apply a residual network ML model to predict the EC on the various scales [43].

The primary objectives of the studies in this category (8 studies) are the prediction of EP, potential for energy savings, and cost-optimal retrofitting solutions [40,43,45,47,49,50,52,59]. As data classification and clustering techniques, some studies (6) adopted K-means [28,32,50,52,57,59]. Ultimately, some (2) applied manual classification [47,49]. As a prediction of EP and cost-optimal retrofit solutions techniques, some approaches (7) employed ANN and GA [40,47,49,50,52,56,57]. Others implemented different ML algorithms, such as random forest (RF) [59]. Lastly, some of the approaches executed simulations and mathematical techniques, such as a multiple linear regression, Pearson's correlation, principal component analysis, Monte Carlo, Gaussian process regression model, Gaussian mixture regression model, and deep learning algorithms [28,49,56]. Finally, some studies (3) use geographical information systems (GIS) and geospatial maps to visualize their results [28,50,56].

5.2. Knowledge Gap

Our analysis concluded that the research gap is related to identifying and testing ML approaches that are best fitted and have better performance in targeting automatic evaluation of buildings' energy performance using EPC data. Moreover, most of the studies use statistical and audit approaches at a multilevel scope [15,17,19,22,24,25,27–41,45,48,49]. However, some studies (13) use the EPC dataset for their analysis [28,29,31,34,50–54,56,57,60,61]. Furthermore, most studies apply simulation techniques and GA for prediction, targeting multi-objective cost-optimal solutions, a promising approach.

We conclude that more research is needed to validate and improve future modeling strategies using EPC datasets and different features. These gaps have shown an opportunity for future research.

5.3. Study Limitations

Although we tried to guarantee the quality of this review and, particularly, the data selection, this study has limitations. Specifically, we would like to highlight the dependency on the keywords and the selected data repositories, since additional data repositories could be used and only English papers were included, neglecting publications written in other languages. Finally, another important limitation of this study is the time frame, given that we focused on papers published in the last five years, between early 2016 and April 2021.

6. Conclusions

The PRISMA methodology summarized the SLR analysis and generated a systematic view of ML and statistical approaches applied in improving the EPB which can be used for future research. This study showed that after 2019, most studies used, processed, and analyzed EPC datasets, adopting ML or statistical approaches. Clustering analysis is applied to find similar patterns in buildings' EPC data. Simulation techniques and K-means clustering are the most used approaches to group buildings with similar characteristics. Box plot statistical analysis and dbscan are robust techniques used to eliminate outliers and noise due to their ability to deal with complex and high-dimensional data. Correlation analysis showed that the best approach is to estimate the importance of each analyzed input dimension. Additionally, the literature indicated that the best and most used evaluation method of the performance of the proposed algorithm was the accuracy of the ML-based solution.

Our research findings aim to fulfill identified knowledge gaps and open a methodological agenda that will help the reader identify effective combinations of ML and statistical approaches, addressing EPB and EERM in the future, providing a good starting point for further research.

Author Contributions: M.A.: Conceptualization, Methodology, Formal analysis, Literature Review, Investigation, Data curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Project administration. V.S.: Conceptualization, Methodology, Formal analysis, Writing—Review and Editing, Supervision, Project administration, Funding acquisition. M.S.D.: Conceptualization, Methodology, Formal analysis, Writing—Review and Editing, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a Ph.D. Scholarship of NOVA IMS supported by project POCI-05-5762-FSE 000223, and its scope lies in the context of Simplex #109 "Consumo SMART". This work is partially funded by national funds through FCT—Foundation for Science and Technology, I.P., under the project FCT UIDB/04466/2020.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank Ricardo Pinto and Vitória Albuquerque [66] for their help reviewing the paper. The authors would also like to thank the reviewers and the editorial team who offered edifying and useful remarks to enhance the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, collection, analysis, or interpretation of data, or writing of the study, nor in the decision to publish the results.

References

1. Rocha, P.; Kaut, M.; Siddiqui, A.S. Energy-efficient building retrofits: An assessment of regulatory proposals under uncertainty. *Energy* **2016**, *101*, 278–287. [CrossRef]
2. A European Green Deal. Available online: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en (accessed on 1 May 2021).
3. Eurostat. *Shedding Light on Energy in the EU—A Guided Tour of Energy Statistics*, 2021st ed.; Eurostat: Luxembourg, 2021. [CrossRef]
4. Canevari, C. *How the EU Built the 2030 Energy Efficiency Target*; European Commission DG Energy: Vienna, Austria, 2018.
5. United Nations. Transforming Our World: The 2030 Agenda for Sustainable Development. 2020. Available online: https://sdgs.un.org/2030agenda (accessed on 13 May 2020).
6. European. Final Energy Consumption by Sector and Fuel in Europe. 2020. Available online: https://www.eea.europa.eu/data-and-maps/indicators/final-energy-consumption-by-sector-10/assessment (accessed on 10 February 2020).

7. Eurostat. Energy Statistics—An Overview. 2021. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview#Final_energy_consumption (accessed on 30 November 2021).
8. Koo, C.; Hong, T. Development of a dynamic operational rating system in energy performance certificates for existing buildings: Geostatistical approach and data-mining technique. *Appl. Energy* **2015**, *154*, 254–270. [CrossRef]
9. Mehmood, M.U.; Chun, D.; Zeeshan; Han, H.; Jeon, G.; Chen, K. A review of the applications of artificial intelligence and big data to buildings for energy-efficiency and a comfortable indoor living environment. *Energy Build.* **2019**, *202*, 109383. [CrossRef]
10. Molina-Solana, M.; Ros, M.; Ruiz, M.D.; Gómez-Romero, J.; Martin-Bautista, M. Data science for building energy management: A review. *Renew. Sustain. Energy Rev.* **2017**, *70*, 598–609. [CrossRef]
11. Vaquero, P. Buildings Energy Certification System in Portugal: Ten years later. *Energy Rep.* **2019**, *6*, 541–547. [CrossRef]
12. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef]
13. VOSviewer—Visualizing Scientific Landscapes. Available online: https://www.vosviewer.com// (accessed on 23 January 2021).
14. Fathi, S.; Srinivasan, R.; Fenner, A.; Fathi, S. Machine learning applications in urban building energy performance forecasting: A systematic review. *Renew. Sustain. Energy Rev.* **2020**, *133*, 110287. [CrossRef]
15. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A review of data-driven approaches for prediction and classification of building energy consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [CrossRef]
16. Wang, Z.; Srinivasan, R. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renew. Sustain. Energy Rev.* **2016**, *75*, 796–808. [CrossRef]
17. Seyedzadeh, S.; Rahimian, F.; Glesk, I.; Roper, M. Machine learning for estimation of building energy consumption and performance: A review. *Vis. Eng.* **2018**, *6*, 5. [CrossRef]
18. Grillone, B.; Danov, S.; Sumper, A.; Cipriano, J.; Mor, G. A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings. *Renew. Sustain. Energy Rev.* **2020**, *131*, 110027. [CrossRef]
19. Pickering, C.; Byrne, J. The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *High. Educ. Res. Dev.* **2014**, *33*, 534–548. [CrossRef]
20. Mendeley, Elsevier. 2019. Available online: https://www.mendeley.com/?interaction_required=true (accessed on 10 January 2017).
21. Semple, S.; Jenkins, D. Variation of energy performance certificate assessments in the European Union. *Energy Policy* **2019**, *137*, 111127. [CrossRef]
22. Scimago Lab. Scimago Journal & Country Rank, Scimago J. Ctry. Rank. 2021. Available online: https://www.scimagojr.com/ (accessed on 10 July 2021).
23. Ascione, F.; Bianco, N.; De Stasio, C.; Mauro, G.M.; Vanoli, G.P. Multi-stage and multi-objective optimization for energy retrofitting a developed hospital reference building: A new approach to assess cost-optimality. *Appl. Energy* **2016**, *174*, 37–68. [CrossRef]
24. Ascione, F.; Bianco, N.; De Stasio, C.; Mauro, G.M.; Vanoli, G.P. A Methodology to Assess and Improve the Impact of Public Energy Policies for Retrofitting the Building Stock: Application to Italian Office Buildings. *Int. J. Heat Technol.* **2016**, *34*, S277–S286. [CrossRef]
25. Ascione, F.; Bianco, N.; De Stasio, C.; Mauro, G.M.; Vanoli, G.P. Addressing Large-Scale Energy Retrofit of a Building Stock via Representative Building Samples: Public and Private Perspectives. *Sustainability* **2017**, *9*, 940. [CrossRef]
26. Ascione, F.; Bianco, N.; De Masi, R.F.; Mauro, G.M.; Vanoli, G.P. Resilience of robust cost-optimal energy retrofit of buildings to global warming: A multi-stage, multi-objective approach. *Energy Build.* **2017**, *153*, 150–167. [CrossRef]
27. Ascione, F.; Bianco, N.; de Masi, R.F.; Mauro, G.M.; Vanoli, G.P. Energy retrofit of educational buildings: Transient energy simulations, model calibration and multi-objective optimisation towards nearly zero-energy performance. *Energy Build* **2017**, *144*, 303–319. [CrossRef]
28. Ali, U.; Shamsi, M.H.; Bohacek, M.; Purcell, K.; Hoare, C.; Mangina, E.; O'Donnell, J. A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making. *Appl. Energy* **2020**, *279*, 115834. [CrossRef]
29. Ali, U.; Shamsi, M.H.; Bohacek, M.; Hoare, C.; Purcell, K.; Mangina, E.; O'Donnell, J. A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings. *Appl. Energy* **2020**, *267*, 114861. [CrossRef]
30. Ali, U.; Shamsi, M.H.; Hoare, C.; Mangina, E.; O'Donnell, J. A data-driven approach for multi-scale building archetypes de-velopment. *Energy Build* **2019**, *202*, 109364. [CrossRef]
31. Gangolells, M.; Gaspar, K.; Casals, M.; Ferré-Bigorra, J.; Forcada, N.; Macarulla, M. Life-cycle environmental and cost-effective energy retrofitting solutions for office stock. *Sustain. Cities Soc.* **2020**, *61*, 102319. [CrossRef]
32. Gangolells, M.; Casals, M.; Ferré-Bigorra, J.; Forcada, N.; Macarulla, M.; Gaspar, K.; Tejedor, B. Office representatives for cost-optimal energy retrofitting analysis: A novel approach using cluster analysis of energy performance certificate databases. *Energy Build.* **2019**, *206*, 109557. [CrossRef]
33. Di Corso, E.; Cerquitelli, T.; Piscitelli, M.S.; Capozzoli, A. Exploring Energy Certificates of Buildings through Unsupervised Data Mining Techniques. *IEEE Xplore* **2017**, 991–998. [CrossRef]
34. Cerquitelli, T.; Di Corso, E.; Proto, S.; Bethaz, P.; Mazzarelli, D.; Capozzoli, A.; Baralis, E.; Mellia, M.; Casagrande, S.; Tamburini, M. A Data-Driven Energy Platform: From Energy Performance Certificates to Human-Readable Knowledge through Dynamic High-Resolution Geospatial Maps. *Electronics* **2020**, *9*, 2132. [CrossRef]

35. Attanasio, A.; Piscitelli, M.S.; Chiusano, S.; Capozzoli, A.; Cerquitelli, T. Towards an Automated, Fast and Interpretable Estimation Model of Heating Energy Demand: A Data-Driven Approach Exploiting Building Energy Certificates. *Energies* **2019**, *12*, 1273. [CrossRef]
36. Rosso, F.; Ciancio, V.; Dell'Olmo, J.; Salata, F. Multi-objective optimization of building retrofit in the Mediterranean climate by means of genetic algorithm application. *Energy Build.* **2020**, *216*, 109945. [CrossRef]
37. Salata, F.; Ciancio, V.; Dell'Olmo, J.; Golasi, I.; Palusci, O.; Coppi, M. Effects of local conditions on the multi-variable and multi-objective energy optimization of residential buildings using genetic algorithms. *Appl. Energy* **2019**, *260*, 114289. [CrossRef]
38. Ruiz, G.R.; Bandera, C.F.; Temes, T.G.-A.; Gutierrez, A.S.-O. Genetic algorithm for building envelope calibration. *Appl. Energy* **2016**, *168*, 691–705. [CrossRef]
39. Niemelä, T.; Kosonen, R.; Jokisalo, J. Cost-effectiveness of energy performance renovation measures in Finnish brick apartment buildings. *Energy Build.* **2017**, *137*, 60–75. [CrossRef]
40. Beccali, M.; Ciulla, G.; Brano, V.L.; Galatioto, A.; Bonomolo, M. Artificial neural network decision support tool for assessment of the energy performance and the refurbishment actions for the non-residential building stock in Southern Italy. *Energy* **2017**, *137*, 1201–1218. [CrossRef]
41. Copiello, S.; Gabrielli, L.; Bonifaci, P. Evaluation of energy retrofit in buildings under conditions of uncertainty: The prominence of the discount rate. *Energy* **2017**, *137*, 104–117. [CrossRef]
42. Bre, F.; Fachinotti, V. A computational multi-objective optimization method to improve energy efficiency and thermal comfort in dwellings. *Energy Build.* **2017**, *154*, 283–294. [CrossRef]
43. Nutkiewicz, A.; Yang, Z.; Jain, R.K. Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Appl. Energy* **2018**, *225*, 1176–1189. [CrossRef]
44. Shen, P.; Braham, W.; Yi, Y. The feasibility and importance of considering climate change impacts in building retrofit analysis. *Appl. Energy* **2018**, *233-234*, 254–270. [CrossRef]
45. Nilashi, M.; Dalvi-Esfahani, M.; Ibrahim, O.; Bagherifard, K.; Mardani, A.; Zakuan, N. A soft computing method for the prediction of energy performance of residential buildings. *Measurement* **2017**, *109*, 268–280. [CrossRef]
46. Eskander, M.; Reyes, M.E.S.; Silva, C.A.S.; Vieira, S.; Sousa, J.M.C. Assessment of energy efficiency measures using multi-objective optimization in Portuguese households. *Sustain. Cities Soc.* **2017**, *35*, 764–773. [CrossRef]
47. Hirvonen, J.; Jokisalo, J.; Heljo, V.J.; Kosonen, R. Towards the EU Emission Targets of 2050: Cost-Effective Emission Reduction in Finnish Detached Houses. *Energies* **2019**, *12*, 4395. [CrossRef]
48. Hosseini, M.; Lee, B.; Vakilinia, S. Energy performance of cool roofs under the impact of actual weather data. *Energy Build.* **2017**, *145*, 284–292. [CrossRef]
49. Fan, Y.; Xia, X. Building retrofit optimization models using notch test data considering energy performance certificate compliance. *Appl. Energy* **2018**, *228*, 2140–2152. [CrossRef]
50. Cecconi, F.R.; Moretti, N.; Tagliabue, L. Application of artificial neutral network and geographic information system to evaluate retrofit potential in public school buildings. *Renew. Sustain. Energy Rev.* **2019**, *110*, 266–277. [CrossRef]
51. Seyedzadeh, S.; Pour Rahimian, F.; Oliver, S.; Rodriguez, S.; Glesk, I. Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making. *Appl. Energy* **2020**, *279*, 115908. [CrossRef]
52. Zekić-Sušac, M.; Scitovski, R.; Has, A. Cluster analysis and artificial neural networks in predicting energy efficiency of public buildings as a cost-saving approach. *Croat. Rev. Econ. Bus. Soc. Stat.* **2018**, *4*, 57–66. [CrossRef]
53. Qu, K.; Chen, X.; Ekambaram, A.; Cui, Y.; Gan, G.; Økland, A.; Riffat, S. A novel holistic EPC related retrofit approach for residential apartment building renovation in Norway. *Sustain. Cities Soc.* **2019**, *54*, 101975. [CrossRef]
54. Ahern, C.; Norton, B. A generalisable bottom-up methodology for deriving a residential stock model from large empirical databases. *Energy Build.* **2020**, *215*, 109886. [CrossRef]
55. Belaïd, F.; Ranjbar, Z.; Massié, C. Exploring the cost-effectiveness of energy efficiency implementation measures in the residential sector. *Energy Policy* **2021**, *150*, 112122. [CrossRef]
56. Von Platten, J.; Sandels, C.; Jörgensson, K.; Karlsson, V.; Mangold, M.; Mjörnell, K. Using Machine Learning to Enrich Building Databases—Methods for Tailored Energy Retrofits. *Energies* **2020**, *13*, 2574. [CrossRef]
57. García-Nieto, P.J.; García-Gonzalo, E.; Paredes-Sánchez, J.P.; Sánchez, A.B. A new hybrid model to foretell thermal power efficiency from energy performance certificates at residential dwellings applying a Gaussian process regression. *Neural Comput. Appl.* **2020**, *33*, 6627–6640. [CrossRef]
58. Geissler, S.; Androutsopoulos, A.; Charalambides, A.G.; Escudero, C.J.; Jensen, O.M.; Kyriacou, O.; Petran, H. ENERFUND—Identifying and rating deep renovation opportunities. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *323*, 012174. [CrossRef]
59. Pistore, L.; Pernigotto, G.; Cappelletti, F.; Gasparella, A.; Romagnoni, P. A stepwise approach integrating feature selection, regression techniques and cluster analysis to identify primary retrofit interventions on large stocks of buildings. *Sustain. Cities Soc.* **2019**, *47*, 101438. [CrossRef]
60. European Commission. Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings. 2002. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=%0ACELEX:32002L0091&from=IT (accessed on 10 March 2019).

61. *Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the Energy Performance of Buildings (Recast)*; European Union: Brussels, Belgium, 2010.
62. EnergyPlus. Available online: https://energyplus.net/ (accessed on 2 March 2019).
63. TRNSYS. Transient System Simulation Tool. Available online: http://www.trnsys.com (accessed on 2 March 2019).
64. Tadeu, S.F.; Alexandre, R.F.; Tadeu, A.J.; Antunes, C.H.; Simões, N.A.; da Silva, P.P. A comparison between cost optimality and return on investment for energy retrofit in buildings—A real options perspective. *Sustain. Cities Soc.* **2016**, *21*, 12–25. [CrossRef]
65. Ciro, G.C.; Dugardin, F.; Yalaoui, F.; Kelly, R. A NSGA-II and NSGA-III comparison for solving an open shop scheduling problem with resource constraints. *IFAC PapersOnLine* **2016**, *49*, 1272–1277. [CrossRef]
66. Albuquerque, V.; Dias, M.S.; Bacao, F. Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 62. [CrossRef]