energies

IMPACT FACTOR 3.2

CITESCORE 7.3

Article

Editor's Choice

AI-Driven Morphological Classification of the Italian School Building Stock: Towards a Deep Energy Renovation Roadmap

Giacomo Caccia, Matteo Cavaglià, Fulvio Re Cecconi, Andrea Giovanni Mainini, Marta Maria Sesana and Elisa Di Giuseppe

MDPI

https://doi.org/10.3390/en18184953

energies

Article

AI-Driven Morphological Classification of the Italian School Building Stock: Towards a Deep Energy Renovation Roadmap

Giacomo Caccia [1], Matteo Cavaglià [1], Fulvio Re Cecconi [1], Andrea Giovanni Mainini [1,*], Marta Maria Sesana [2] and Elisa Di Giuseppe [3]

[1] Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Via Ponzio 31, 20133 Milan, Italy; giacomo.caccia@polimi.it (G.C.); matteo.cavaglia@polimi.it (M.C.); fulvio.rececconi@polimi.it (F.R.C.)
[2] Department of Civil, Environmental, Architectural Engineering and Mathematics, University of Brescia, 25123 Brescia, Italy; marta.sesana@unibs.it
[3] Department of Construction, Civil Engineering and Architecture (DICEA), Università Politecnica delle Marche, 60131 Ancona, Italy; e.digiuseppe@staff.univpm.it
* Correspondence: andreagiovanni.mainini@polimi.it; Tel.: +39-02-23996018

Abstract

The Italian school building stock is largely outdated, with structural and technological inadequacies leading to low comfort and high energy consumption. Addressing this challenge requires large-scale renovation supported by an integrated, data-driven approach. This study conducted a nationwide analysis of over 40,000 school buildings. After incomplete or inconsistent records were filtered out, a refined subset was selected. Building forms were reconstructed by cross-referencing GIS data with multiple open data sources. Using supervised machine learning, the research identifies and classifies recurring morphological patterns to define a set of 3D school building archetypes. These archetypes are enriched with spatial configurations and physical characteristics aligned with national educational standards. The result is a macrotypological classification based on form, conceived as part of an operational tool to support policymakers, designers, and public administrations in selecting effective retrofit strategies. This contributes to the creation of large-scale national renovation strategies, as well as Renovation Roadmaps and Digital Building Logbooks in line with the Energy Performance of Buildings Directive (EPBD IV), specifically tailored to the Italian context. The novelty of this work lies in its unprecedented scale and the use of AI to enable fast, replicable assessments of retrofit potential, thereby supporting informed decisions in energy-efficient renovation planning.

Keywords: energy performance; schools and educational buildings; GIS; synthetic images; data augmentation; cluster analysis; *k*-nearest neighbors (kNN); architectural engineering; form factor; building renovation passport

Academic Editor: Christian Inard

Received: 1 August 2025
Revised: 7 September 2025
Accepted: 13 September 2025
Published: 17 September 2025

Citation: Caccia, G.; Cavaglià, M.; Re Cecconi, F.; Mainini, A.G.; Sesana, M.M.; Di Giuseppe, E. AI-Driven Morphological Classification of the Italian School Building Stock: Towards a Deep Energy Renovation Roadmap. *Energies* **2025**, *18*, 4953. https://doi.org/10.3390/en18184953

1. Introduction

The educational building stock in Italy represents a critical sector for achieving national decarbonization goals. A significant portion of school buildings, predominantly constructed after 1975 (Figure 1), are now obsolete and energy-inefficient [1]. Technological deficiencies in the building envelope, such as poor insulation and low-performance windows, as well as the inefficiency of HVAC and lighting systems, contribute to high energy consumption and, most importantly, fail to guarantee adequate thermohygrometric, visual, and acoustic comfort conditions necessary for a healthy and productive learning environment.

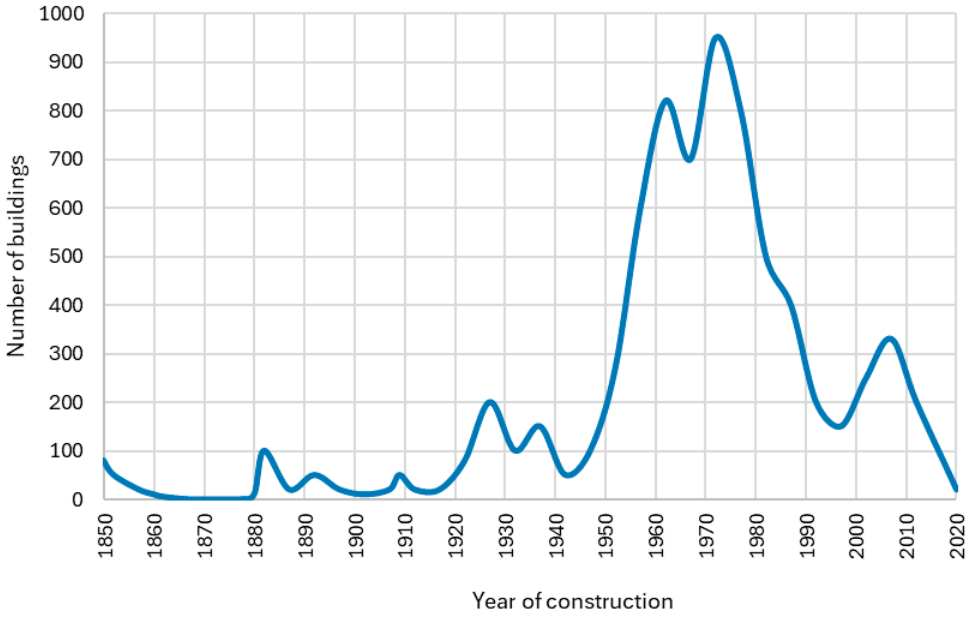


Figure 1. Distribution of the number of constructed school buildings over time.

Some initiatives, like the RERoadS project [2], are actively addressing the urgent need for large-scale deep renovation in public buildings such as schools by developing integrated approaches that combine rapid and easily applicable forecasting tools and data-driven methodologies, including the use of machine learning to process existing data and derive morphological, technical, and construction information useful for building simulation and analysis. In line with the recast Energy Performance of Buildings Directive (EPBD) IV [3], RERoadS also focuses on the definition of Digital Building Logbooks and structured Renovation Roadmaps, supporting the future implementation of Building Renovation Passports specifically tailored to the Italian school stock. In this context, the presented process serves as a useful starting point for achieving the RERoads research goals.

An energy retrofitting process cannot be undertaken without a thorough understanding of the building's current condition. However, analyses conducted on a national scale require an approach that moves beyond the individual case study. Many studies [4–6] have demonstrated that a building's energy performance is intrinsically linked not only to materials and systems but also to its shape. Despite this awareness, the scientific literature still lacks studies that attempt a systematic classification of school building types based on their form and a consequent assessment of energy implications. Analyses are often limited to a few representative archetypes, thus neglecting the real heterogeneity of the existing stock.

This paper addresses this gap to overcome these limitations. The research proposes an innovative methodology to analyze and classify the entire Italian school building stock, using a dataset of over 40,000 buildings extracted from the Open Data platform of the Italian *Ministero dell'Istruzione e del Merito* [7]. Unlike previous studies focused on smaller samples [8–11], this work leverages the computational power and pattern-recognition capabilities of machine learning algorithms to analyze a dataset of this magnitude. Supervised learning techniques were employed to identify the most recurrent school building shapes and define reference educational buildings, based on the typical forms emerging from a systematic literature review aimed at establishing a robust morphological taxonomy. The novelty of this approach is twofold. First, the scale of the dataset provides a comprehensive and statistically significant overview of the built environment. Second, the use of machine learning offers a powerful and scalable tool for analyzing the morphology of complex building stocks.

The outcome of this research is a classification of the Italian school building stock into macrotypological categories, each characterized by a specific, predefined form. Furthermore, the framework enables the development of three-dimensional archetypes of school buildings, integrating information from both the database analysis and the geocoding and classification process. Importantly, this classification is not intended as the conclusion of the research but, rather, as an operational tool to support policymakers, designers, and public administrations, providing strategic guidance for the selection of the most suitable and effective large-scale energy-retrofit interventions.

2. Background

The Italian climate is characterized by six climatic areas, from A, the hottest, to F, the coldest, determined based on heating degree days (HDDs, in Italian *Gradi Giorno*—GG), established through D.P.R. 412/1993 [12].

The Italian school system, defined by *Decreto Ministeriale 18 dicembre 1975, n. 29* [13] (DM 75), is composed of four education grades (kindergarten, primary school, middle school, and high school), which involve students from 3 to 18 years old. Detailed information about Italian climatic zones and Italian school grades is provided in Appendix A. In addition to the school grades definition, DM 75 is a fundamental and innovative standard for school buildings, aiming to ensure safe, functional, and educationally adequate environments. It established requirements for the size and configuration of classrooms, the inclusion of laboratories and ancillary spaces, and the flexibility to adapt learning environments to different teaching needs. Emphasis was placed on the outdoor dimension, requiring schools to be located on healthy sites, away from pollution and traffic, and equipped with green areas, with a limit of one-third of the site permitted for construction.

Beyond technical aspects, DM 75 recognized schools as civic infrastructures, integrated into local planning frameworks and responsive to demographic and social contexts. Although formally suspended in 1996, several parameters, most notably the space-per-student ratios (1.80 m^2 for primary and 1.96 m^2 for secondary schools), remain widely adopted in practice. Due to the high number of educational facilities built in this period, the standard remains highly relevant for large-scale analyses and has a particular relevance also for broad analyses on the Italian school building stock.

2.1. Relationship Between Form Factors and Energy Needs

Recent research increasingly highlights how morphological characteristics, such as surface-to-volume ratios (S/V), compactness, and orientation, play a key role in determining both heating and cooling demands across various climatic conditions.

The relationship between compactness and heating demand is emphasized by Marincu et al. [14], who investigated the trade-off between wall insulation thickness and the surface-to-volume ratio. Their results reveal that more compact forms (lower S/V ratios) significantly reduce thermal bridging and heating demand, with high S/V ratios leading to exponential increases in thermal transfer losses. Achieving passive house standards was found to be viable for buildings with low S/V ratios using moderate insulation levels, whereas highly articulated forms have more difficulty meeting the same targets, even with extensive insulation.

Baglivo et al. [15] explored the influence of the S/V ratio in the context of climate change adaptation. Their simulations, projected up to 2070, indicate that compact forms are less sensitive to climate-induced variations in energy demand. Buildings with low S/V ratios demonstrated greater stability in energy performance across future climate scenarios. Conversely, highly articulated shapes, especially in warmer Italian regions, exhibited up to 20% increases in energy demand over time. The authors argue that over-reliance on

insulation alone is neither cost-effective nor sufficient, advocating instead for a holistic approach that integrates form, thermal properties, and urban design.

In addition, in warm climates such as those of Athens and Seville, Premrov et al. [1] demonstrated that the aspect ratio and the vertical and horizontal extensions of timber–glass buildings have a significant impact on cooling energy demand. Buildings with south-facing glazing experience an exponential increase in cooling loads as the aspect ratio grows, whereas north-facing orientations are less affected. These results highlight the critical role of building shape in mitigating overheating risks, especially for lightweight constructions.

From a design methodology perspective, Parasonis et al. [6] introduced the concept of geometric efficiency, defined as the ratio between external envelope area and usable internal space, as a tool for early-stage optimization. Their findings demonstrated that building geometry significantly impacts both operational energy use and material requirements, particularly for smaller buildings. Inefficiencies in form can, to some extent, be offset by improved envelope thermal properties, though this becomes more challenging in compact or single-family typologies.

Torabi et al. [16] focus their research on identifying which design parameters most significantly impact embodied and operational carbon and Energy Use Intensity (EUI), pursuing a multi-objective optimization study. The research highlights how geometry parameters, especially building shape, are highly influential on energy performance in all the considered cases: compact shapes, particularly square ones, emerged as the most beneficial in terms of EUI.

Kistelegdi et al. [17] provided a systematic classification of building geometry design variables used in energy and comfort optimization studies. Their review reveals that most studies rely on simplified, modifiable geometric descriptors such as aspect ratios or depth, with a limited exploration of more detailed spatial or morphological variables. They also note that early-stage geometric optimization offers a higher return on investment compared to later-stage improvements in materials or HVAC systems.

Finally, Li et al. [18] called attention to the regional sensitivity of the S/V ratio and its limitations as a stand-alone metric. Their review advocates for revised S/V ratio definitions incorporating solar radiation, material properties, and orientation. They report that, in cold regions, lower S/V ratio values correlate strongly with reduced heating energy use, while in tropical zones, the impact is more pronounced on cooling loads.

Despite these advances, the classification of forms in relation to building energy performance remains quite underexplored, particularly for school buildings, which represent a significant portion of the public building stock, and where a significant portion of human life is spent.

2.2. State of the Art of Classification and Clustering Methods of School Buildings

As a first step, a literature review was conducted to gain a global overview of school reference building identification. An analysis of the state of the art, in line with the PRISMA framework [19], was performed using Scopus as the main database for scientific literature, with the following query:

KEY ((educational AND building* OR school AND building*) AND (classification OR typology OR categorization OR cluster*)),

which allowed for a fast and efficient identification of the state of the art of the research focused on advancing the education field through the development of clustering and identification techniques for school building typologies, aimed at supporting building energy simulations. The analysis initially yielded 74 papers, with an additional 15 publications identified through other sources and citation searching to broaden the scope of the review. The following criteria were subsequently applied to refine and focus the literature review.

- Only English-language contributions were considered.
- Only open-access contributions were considered.
- Only topic-related contributions were included (after title and abstract reading).

Ultimately, a total of 27 papers were considered. The first outcome of this analysis is presented in Table 1, which summarizes the methodologies used to classify schools. Additionally, to better understand the typical school building samples used in current research, the number of schools considered in each study was also collected and included in the table.

Table 1. Classification method used for the identification of school building clusters.

Ref.	Year	Location	Sample	Classification Method
RSE [20]	2024	Italy	-	One reference building for each school type
Pedone et al. [11]	2023	Foggia, Italy	81	No clusters, schools modeled individually
Campagna & Fiorito [21]	2023	Puglia, Italy	1090	K-means clustering based on S/V ratio and year of construction
Kazem & Al-Kazzaz [22]	2023	Iraq	-	Clusters made on five pre-defined building shapes, number of classes, and school type (primary, secondary, …)
Bo and De Angelis [23]	2022	Milan Municipality, Italy	277	No clusters, schools modeled individually, UBEM approach
Geraldi et al. [24]	2021	Brazil	298	Organized in seven clusters based on pre-defined shapes
Zinzi et al. [25]	2021	Italy	-	Simulation made on one representative reference building
Alghamdi et al. [8]	2020	University of British Columbia, Canada	71	Fuzzy clustering on water, energy, and carbon flows' real data records
Cukovic Ignjatovic [26]	2020	Serbia	563	TABULA-like approach
Akil et al. [27]	2019	France, Pas-de-Calais	117	k-means clustering based on water, gas, and electricity real data records
Marrone et al. [10]	2018	Lazio, Italy	60	K-means clustering
Zhang et al. [28]	2017	China, cold climate areas	170 schools, 207 buildings	Organized in eight clusters based on pre-defined shapes
Salvalai et al. [9]	2017	Lecco, Italy	38	Clusters made on four pre-defined building shapes, number of floors, and % of glass surface
Haj Hussein et al. [29]	2016	Palestina	-	Organized in six clusters based on pre-defined shapes
Arambula et al. [30]	2015	Treviso Province, Italy	60	K-means clustering
Morck et al. [31]	2015	German, Denmark, Italy and Norway	-	Simulation made on three representative reference buildings
Santamouris et al. [32]	2007	Greece	320	Fuzzy clustering

The table outlines 17 studies conducted between 2007 and 2024 on the classification and clustering of school buildings across various geographic contexts. Among these, nine studies, more than half of the considered studies, explicitly organize buildings into clusters based on predefined shapes or form-related features. This indicates a methodological trend through which geometric configuration is treated as a key classifier.

For example, Geraldi et al. [24] clustered 298 Brazilian schools into seven groups based on predefined shapes, and Zhang et al. [28] did similarly for 207 buildings in cold regions of China, forming eight shape-based clusters. Likewise, Kazem & Al-Kazzaz [22] categorized Iraqi schools into five predefined shapes, in combination with the number of classes and school type (e.g., primary or secondary).

Additionally, Campagna & Fiorito [21] applied k-means clustering to a notably large dataset (1090 buildings in Puglia, Italy), using a combination of shape-sensitive indicators

such as the surface-to-volume ratio (S/V) and year of construction, reinforcing the central role of form in large-scale typological classifications.

On the other hand, five studies used alternative clustering strategies based on operational or environmental performance data, indicating a more dynamic and data-driven approach. For instance, Alghamdi et al. [8] used fuzzy clustering for 71 buildings at the University of British Columbia, grouping them based on actual water, energy, and carbon flow records. Akil et al. [27] analyzed 117 buildings in northern France with k-means clustering based on gas, water, and electricity usage. Similarly, Santamouris et al. [32] used fuzzy clustering for 320 schools in Greece, one of the earliest applications of such methods in this domain.

The TABULA approach used by Cukovic Ignjatovic [26] for 563 Serbian schools represents another form of classification based on typologies derived from European building stock characteristics. This method organizes buildings according to standardized typological templates, often including construction period, geometry, and envelope performance.

Three studies, those of Pedone et al. [11], Bo & De Angelis [23], and Kazem and Al-Kazzaz [22], explicitly modeled buildings individually, with the choice not to apply clustering methods. While this approach enables detailed, building-specific simulations, it inherently limits scalability when applied to large datasets. The other three studies, by Zinzi et al. [25] and Morck et al. [31], as well as the recent RSE report [20], relied on representative reference buildings, which serve as typological exemplars rather than statistical clusters.

As the shape-based approach has emerged as the most recurrent and has been adopted in recent research, a focused examination of this method is now presented. A detailed analysis was conducted for the Italian context, based on the *Manuale di edilizia scolastica* by M. Sole [19] and *L'architettura degli edifici per l'istruzione* by A. di Bitonto and F. Giordano [20], which provide significant insights into the Italian school building stock. These works show how school environments reflect the educational and cultural paradigms of their time, demonstrating a close link between building morphology and pedagogical models. Evolving teaching methods were progressively translated into distinct architectural layouts, ultimately resulting in the development of specific school typologies.

Until the Industrial Revolution, educational facilities lacked a dedicated architectural form, often reusing typologically similar structures—such as barracks, seminaries, or monasteries. These buildings were inward-focused, arranged around a central cloister, and functioned as contemplative, secluded spaces.

In the 19th century, the German "block school building" model became dominant in Italy, characterized by a corridor-based distribution. From this, two main typologies emerged:

- Linear block: Classrooms aligned along a corridor, usually oriented northward to maximize daylight. Corridors also served as social spaces. Classrooms ranged from 55 to 80 m^2 with ceiling heights between 4 and 4.5 m. Floor plans adopted regular "I," "L," or "C" shapes, with the long side hosting the corridor-classroom system and short sides for auxiliary functions.
- Block with internal courtyard: similar to the linear block but arranged around an internal courtyard, resulting in an "O"-shaped plan.

Subsequent variations included the following:

- Linear block with internal void: classrooms accessed via balconies around a central void, with classrooms on both sides.
- Outdoor or extensive school: In contrast, this typology emphasized outdoor connectivity and horizontal development. Common layouts include cross-shaped wings extending from a central core, or comb patterns with functional wings (e.g., gyms,

labs, canteens) branching from a linear spine. Floors took the form of "E," "T," or "F" shapes, with outdoor areas playing a central role.

Following post-war reconstruction, socio-economic growth and educational reforms spurred architectural innovation. The corridor model was replaced with "functional units," transforming school space organization. Large, purpose-driven cores replaced hallways, fostering diverse learning settings centered on student engagement.

This pedagogical evolution shifted the teacher–student relationship, positioning students as active participants. Traditional layouts gave way to ateliers, labs, learning gardens, and other varied environments. By the 1960s, frontal teaching methods were widely abandoned in favor of the "pavilion" school typology, which continues to influence current designs. Pavilion schools introduced double-sided lighting, flexible layouts, generous communal areas, eliminated corridors, shortened hallways, and decentralized service zones.

The analysis of the most recurrent building shapes was expanded to include studies identified through the PRISMA analysis, which encompassed not only Italian research but also international and foreign studies. The results are presented in Table 2, which outlines the various school building shapes identified across the selected studies. To enhance clarity and readability, some data from the original studies were reinterpreted during the creation of the table: U-shaped and C-shaped buildings were grouped under a single 'C' category, and an 'Other' category was introduced to consolidate all undefined or uncommon shapes described in the literature.

Table 2. Identified shapes in topic-related studies. The X indicates that a specific school shape or type is considered in the analyzed research.

Ref.	Year	Study Area	E	H	L	O	C	R	High Rise	Stepped	Other	Notes
RSE [20]	2024	Italy				X	X	X			X	
Kazem & Al-Kazzaz [22]	2023	Iraq			X	X	X	X			X	
Geraldi et al. [24]	2021	Brasil	X	X	X	X	X	X			X	
Zinzi et al. [25]	2021	Italy						X				Exposed corridor
Elitsa Ivanova [33]	2019	-	X	X	X	X	X	X			X	Grouped in six macro-categories: courtyard, block, cluster, and town-like
Zhang et al. [28]	2017	China		X	X		X	X	X		X	Distinction between H with and without atrium
Salvalai et al., 2017 [9]	2017	Lecco, Italy			X	X	X	X		X		
Haj Hussein et al., 2016 [29]	2016	Palestine			X	X	X	X				Distinction between O with a covered patio or not
Morck et al. [31]	2015	Germany Denmark Italy Norway				X		X				Distinction between R with corridor exposed and not
Rigolon [34]	2010	Europe	X	X	X	X					X	Grouped in four macro-categories: courtyard, block, cluster, and town-like
M. Sole [35]	1995	Italy	X		X	X	X	X			X	Elaboration and synthesis of the two sources
di Bitonto & Giordano [36]	1995	Italy										

This comparison reveals significant heterogeneity in the classification of school building forms, with no shared framework or typological consensus across the reviewed literature. The most frequently recurring shapes are L, C, and H, which appear in at least seven of the analyzed sources. Nevertheless, the interpretation and inclusion of these types vary markedly across authors, suggesting that typological decisions are primarily driven by subjective criteria, rather than standardized definitions or objective classification protocols.

For instance, Geraldi et al. [24] and Ivanova [33] both include a broad range of forms, but they group these within different macro-categories; similarly, Rigolon [34] refers to four macro-categories (courtyard, block, cluster, and town-like) but provides limited methodological details regarding the distinction between types.

The category "Other" collects different shapes, such as the multilinear one identified by Ivanova [33], composed of a main rectangular core and several smaller branches, similar to E-shape but with more branches and placed in more irregular positions or the multiple typology, consisting of buildings with different shapes connected, identified for high school by RSE [20] and Geraldi [24].

Considering other building uses outside Italy, Dahlström et al. [37] analysed the Swedish residential stock, focusing on single-family and multi-family houses in Uppsala and Gotland. The authors applied the k-means algorithm simultaneously considering ten parameters, both technical and socio-economic. Their results identified twelve archetypes in Uppsala and thirteen in Gotland, showing that combining technical and socio-economic parameters provides a robust segmentation of modern residential stocks for hybrid UBEM applications.

Lucchi et al. [38] applied a density-based clustering method (HDBSCAN) to the historical town of Calavino, Italy. The dataset included residential, commercial, abandoned, and underused buildings, with particular attention to cultural heritage constraints. Both physical and morphological variables—such as building surface, number of floors, height, surface-to-volume ratio, and typological form (aggregated, isolated, linear)—were explicitly considered, alongside functional and conservation parameters. The analysis revealed that clustering based solely on physical attributes resulted in fewer clusters and outliers, whereas incorporating functional and heritage-related variables produced a much larger number of outliers. This finding underscores the high variability of historic urban fabrics and the challenge of grouping atypical buildings into homogeneous categories.

In conclusion, while certain building forms recur throughout the literature, there is no unified or standardized framework for classifying school building types. Each study tends to adopt its own interpretive model, often shaped by geographic, cultural, or functional contexts, and seldom supported by a robust methodological foundation. This lack of consistent background analysis significantly hinders cross-comparability between studies. To address these limitations, the following chapter introduces a data-driven approach based on the entire Italian building stock, aiming to identify representative school building typologies and overcome the fragmented, author-dependent classification landscape revealed in the literature review.

3. Methodology

This section outlines the methodological workflow adopted to achieve a large-scale classification of the Italian school building stock. The approach integrates national open datasets with advanced geospatial analysis and supervised machine learning techniques to identify and cluster recurrent morphological patterns. The following subsections detail the dataset preparation and cleaning process, the extraction and classification of building footprints, and the generation of a comprehensive reference dataset structured by school type, climate zone, and morphological features. A schematic representation of the research workflow and outputs is reported in Figure 2.

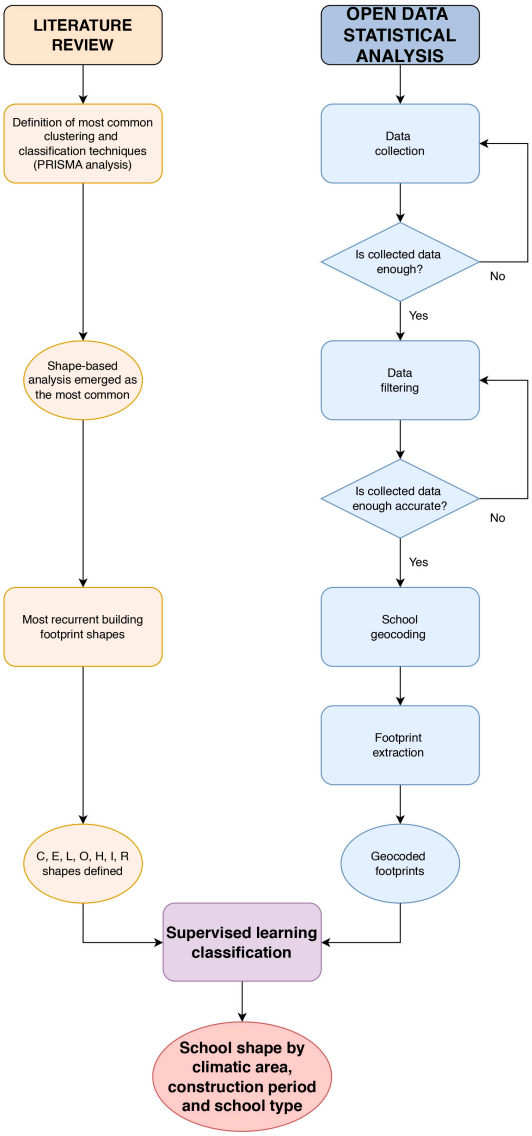


Figure 2. Methodological workflow for Italian school shape definition. Rectangles: methodological steps. Ovals: intermediate and final findings/results.

3.1. Dataset Description and Cleaning

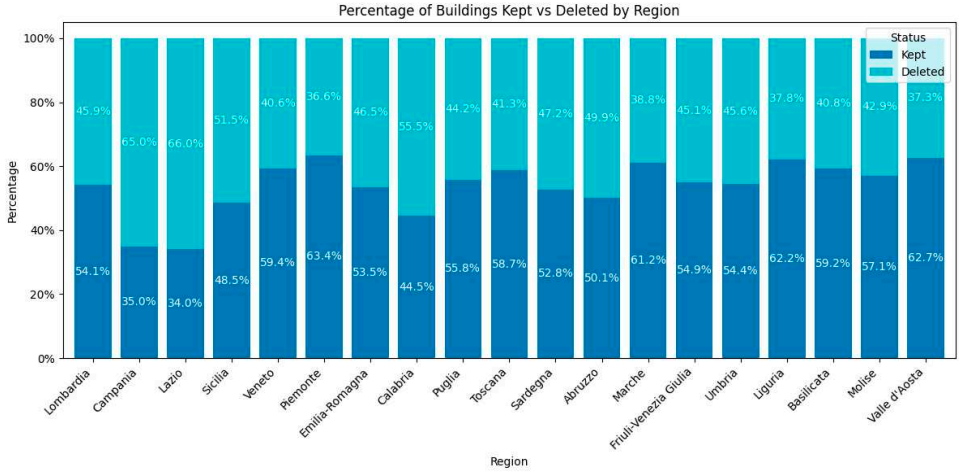
A comprehensive dataset containing all the available data on Italian schools was created by merging the files from the *Ministero dell'Istruzione e del Merito* OpenData platform [7]. The dataset consists of 61,307 records, characterized by a school code (*codicescuola*), representing the type of school complex, and the building code (*codiceedificio*), identifying the physical building. The dataset includes only public schools for all the Italian regions except Trentino-Alto Adige.

With a consideration of the entire dataset, which consists of 48,988 school complexes located in 40,133 school buildings, some inconsistencies and incorrect data were identified. This database provides a very detailed collection of information about school buildings; however, some entries are difficult to calculate, measure, or correctly input into the platform by the designated technician. Therefore, a cleaning rule was applied, excluding only those records in which at least one of the three geometry columns (*volumelordoedificio*, *superficietotalearealibera*, and *superficietotaleareascolastica*) outside the 5th and 95th percentiles has been introduced. After this first level of data cleaning, the total number of records was

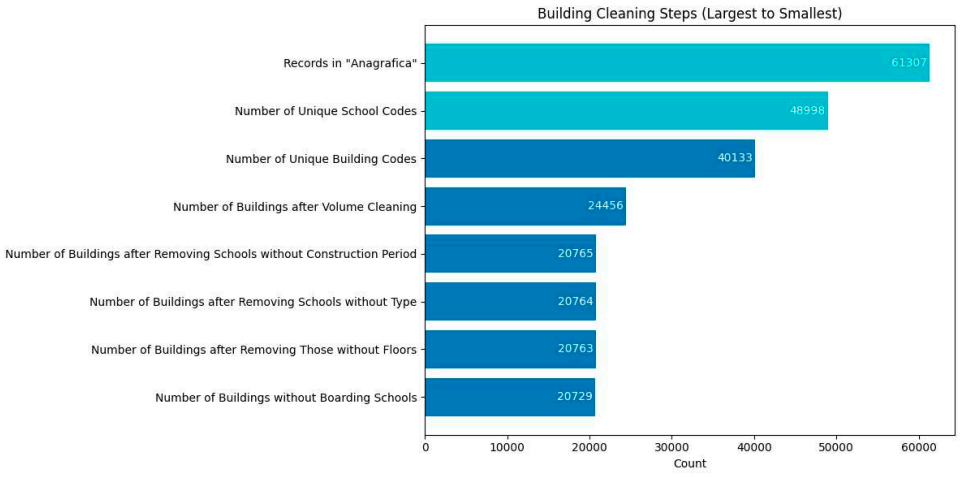
reduced to 24,456. Second-level filtering was then applied, further cleaning and processing the data. This second phase included the exclusion of the following:

- Records with "Not defined" values in either the year of construction (*annocostruzione*) or construction period (*periodocostruzione*) columns, reducing the total number of records to 20,765;
- Records with zero floors and no intermediate floor height;
- Records classified as "*convitto nazionale*", as these include residential functions within the current research framework.

The total number of records in the final dataset was 20,729 buildings, corresponding to 51.6% of the total. A summary breakdown of the cleaning process by region and the number of schools considered after every cleaning step is shown in Figure 3.



(**a**)



(**b**)

Figure 3. Percentage of retained vs. deleted buildings by region (**a**). Number of retained schools after each cleaning step (**b**).

An analysis of the final dataset reveals that most of the excluded buildings are located in the regions of Lazio and Campania, which account for approximately 65% of the removals. In contrast, the region with the lowest exclusion rate was Valle d'Aosta, with only 37.3% of buildings discarded. On average, about 50% of the records were removed across all regions.

3.2. Footprint Extraction, Validation, and Supervised Learning Classification

Owing to its extensive coverage across the Italian territory and open accessibility, the Microsoft Building Footprints (MBF) dataset was selected to extract the building footprints corresponding to the filtered school dataset. The addresses obtained from the Open Data platform were first geocoded using OpenStreetMap (OSM) and then matched to the nearest building footprint in the MBF dataset. The matching process was based on the distance between the geocoded point and the centroid of each building footprint, with a maximum threshold of 200 m considered a valid association. This procedure resulted in a new geocoded dataset containing 14,478 buildings with successfully associated footprints, corresponding to a success rate of 59.2%. Overall, this dataset represents approximately 36% of the entire Italian school building stock.

To validate the geocoding results and improve the reliability of the sample, the identified building footprints were cross-referenced with those available in OpenStreetMap (OSM) via QuickOSM QGIS plug-in [39]. This comparison allowed the creation of a refined subset of school footprints used in the identification of typical plan shapes. An overview of this process is provided in Figure 4.

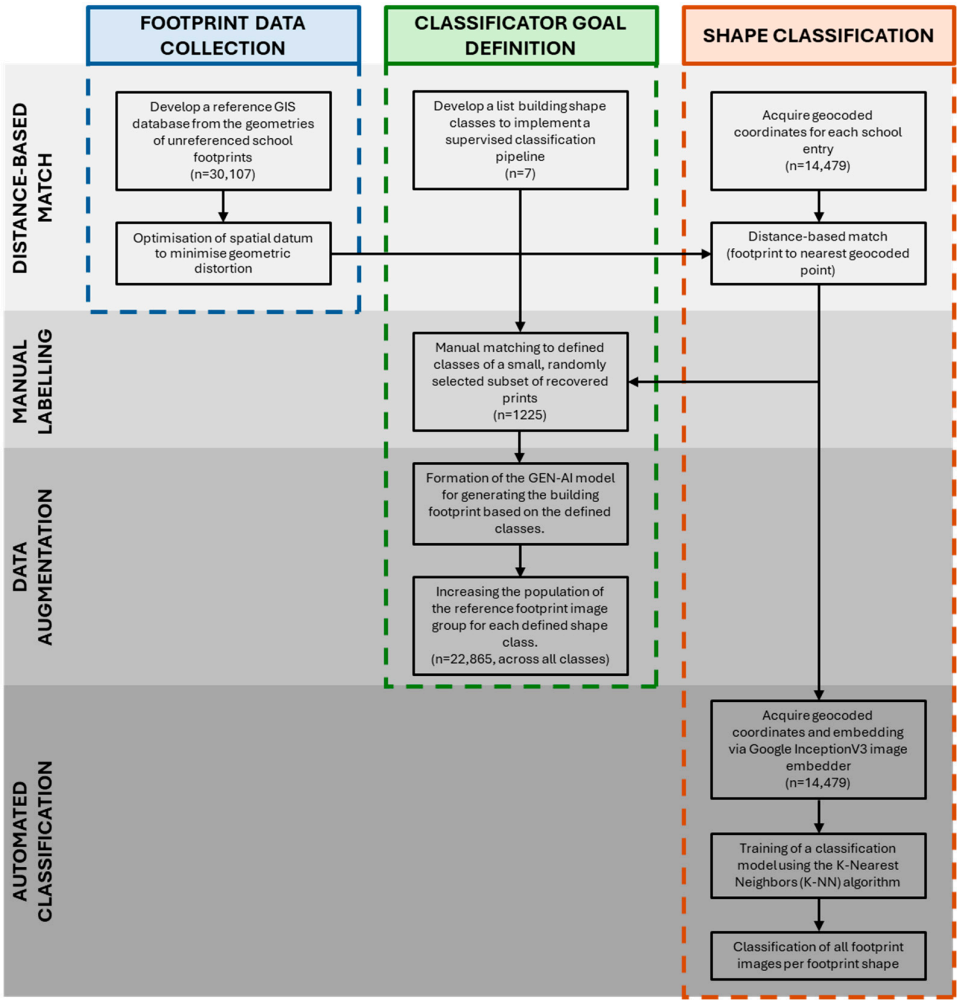


Figure 4. Summary of the shape classification process.

Since a full 3D analysis would be too detailed and resource-intensive for the scope of this study, the 2D building footprint was used as a practical proxy to capture key spatial and morphological characteristics. To this end, a classification pipeline was developed to assign each school to one of seven predefined plan-shape categories (I, L, O, E, H,

C, and R, where R stands for "regular and compact" shapes), as shown in Figure 5. The pipeline integrates automated matching of geocoded school coordinates with corresponding building footprints, followed by an image-based classifier applied to rasterized versions of those footprints.

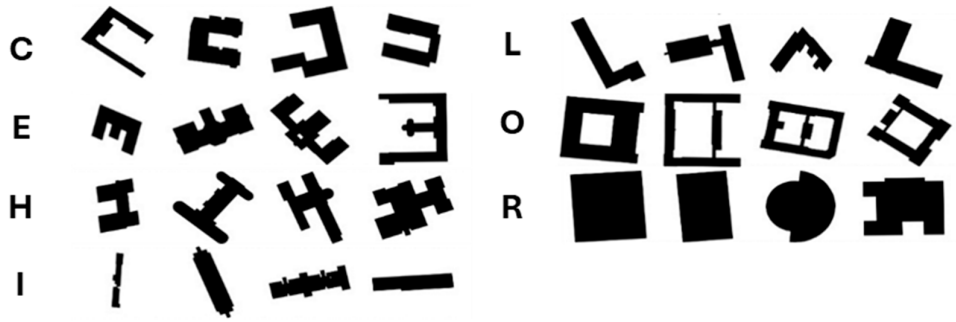


Figure 5. Predefined shapes alongside some related examples from the Italian school stock.

Because the original school dataset only provided postal addresses, another geocoding process was first applied to derive precise latitude–longitude coordinates. These geocoded points served as reference anchors for proximity-based matching routines that linked each school to the nearest building footprint in the nationwide reference dataset. This proximity-based approach is standard in GIS spatial analysis, particularly when working with sparse point data lacking detailed spatial context.

The reference database of building footprints was constructed from GIS data by querying the OpenStreetMap (OSM) database at the national scale to retrieve all building outlines and, when available, the perimeter of school complex amenities. The extracted geometries, originally projected in the WGS84 datum, were subsequently reprojected to the RDN2008 coordinate system to minimize shape distortion during spatial analysis. This step was crucial, as the footprints are later classified based on their morphological shape. A representation of the final Italy project and the position of school complexes and geocoded addresses is shown in Figure 6.

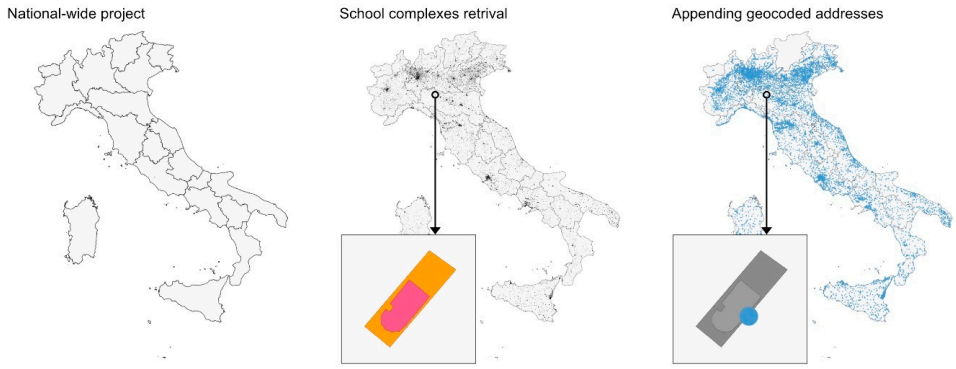


Figure 6. From left to right: Italy project, school complex locations from OSM, and geocoded address positions.

Distance-based matching was conducted using a fidelity threshold of 300 m, with matches considered valid only when the distance between the geocoded school address and the candidate footprint was below this limit. The 300 m threshold, while seemingly large, was adopted to accommodate cases where geocoded anchors, most often street access points, are positioned at a considerable distance from the actual school building, such as when the latter is situated within a courtyard or an enclosed lot. Distances between the geocoded anchors and the corresponding school buildings were analyzed, revealing a

median matching distance of 60 to 80 m (Figure 7). This indicates that the selected threshold was conservative yet reliable, ensuring that most schools were correctly matched while accounting for occasional cases where access points are located further away from the building itself. Unlike standard methods that use centroid-to-point distance, this analysis measured the shortest distance from the anchor point to the perimeter of each footprint (Figure 7), improving reliability in dense areas with multiple nearby buildings [40]. To avoid duplicate assignments, matched footprints were removed from the candidate pool after each successful match. Using this approach, 29% of unique schools in one building (n = 4151) were successfully linked to a corresponding building footprint.

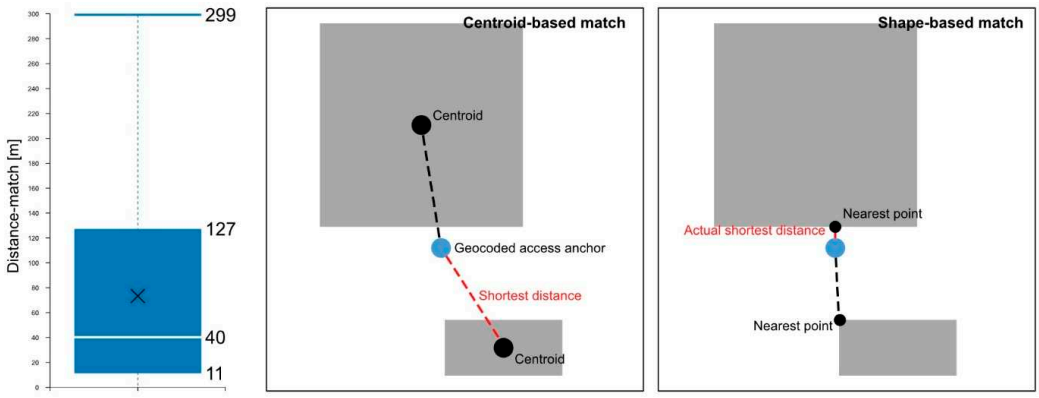


Figure 7. Distance calculation method. The X in the boxplot chart indicates the average distance match.

All matched school buildings were exported as 256 × 256 pixel square images. A randomly selected subset was manually labeled into the seven predefined shape categories to create a reference dataset for training a supervised classification model (n = 1225, average per class = 175), allowing the model to learn to categorize new inputs based on labeled examples.

A practical example of the footprint extraction methodology is provided in Figure 8. The process begins with the coordinates of the point extracted with OSM from the Open Data Database, which are later used to extract the validated footprints with the QuickOSM tool. As shown in the figure, the first geocoding from Open Data was not accurate, as the point does not indicate the exact location of the school. However, the adopted checking process correctly identified the school (the Kindergarten Carlo Collodi, Lughignano, TV, Italy, in this case), considering the nearest one within a range of 300 m.

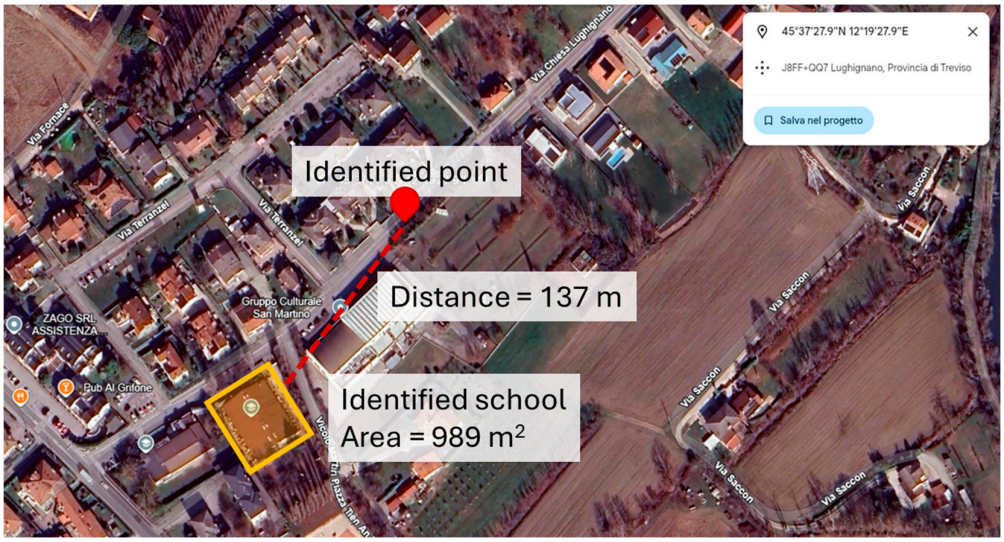


Figure 8. Verification of geocoding and footprint extraction framework.

The building footprint association process achieved an average success rate of 40% across regions, ranging from a minimum of 20% in Umbria to a maximum of 64% in Veneto. Regional representativeness is shown in Figure 9, where the percentage of data kept for the construction of the database for image classification is shown step by step. Overall, each region is adequately represented in the final dataset, maintaining a relatively constant proportion at each step. The regions most affected by data reduction were Molise, Valle d'Aosta, and Calabria. Although data loss may appear significant, it is considered acceptable when accounting for the intrinsic limitations of the public Open Data dataset and the user-generated, incomplete, and non-uniform compilation of the OSM database across Italy. Conversely, excluding uncertain and unverifiable records enabled the construction of robust archetypes that accurately represent school buildings and their morphological characteristics.

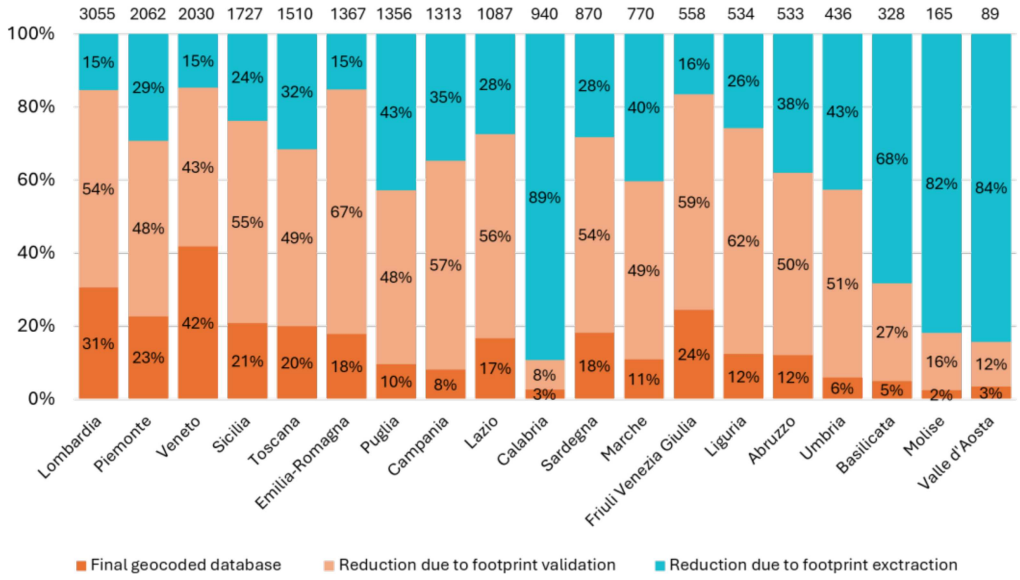


Figure 9. Comparison between the dataset after the data cleaning and the final one used for shape classification, following geocoding, footprint extraction, and validation. Regions on the *x*-axis are ordered according to the number of schools in the cleaned dataset.

K-nearest neighbors (kNN) was selected as the method for automated classification due to its simplicity and flexibility in image-based tasks [41–43]. To address the challenges posed by high-dimensional data and to enhance classification accuracy, each image was first processed through a pre-trained embedding model, specifically, Google Inception V3. This deep convolutional neural network extracts compact and meaningful feature representations from input images, reducing dimensionality and allowing the kNN algorithm to operate on more informative and manageable data for shape classification [44].

Because the kNN algorithm relies on comparing new input data with previously stored feature representations, enhancing the diversity of the reference dataset was essential to improve classification performance. In the case of building footprints, which lack a consistent orientation rule, a dedicated data augmentation process was implemented to expand the representativeness of the training set (Figure 10). This process addressed two main objectives: (1) increasing the morphological variety of building shapes by incorporating additional footprint samples with different geometries and (2) generating multiple rotated instances of each footprint to account for orientation variability within the image plane. This approach ensures that the classifier can robustly recognize and categorize building shapes regardless of their angular alignment or design complexity. Goal (1) was achieved by training class-specific image generation LoRA models using the Stable Diffusion (SD)

platform, enabling the generation of synthetic building footprints that comply with the predefined shape categories [45]. Each LoRA model was trained following the Stable Diffusion 1.5 LoRA training specifications available in the OneTrainer suite [46], using the manually labeled dataset (n = 1225), one class at a time. The generated images were subsequently manually validated, and only those consistent with the intended class were added to the augmented training set. Goal (2) was addressed by augmenting each reference image through 16 incremental rotation steps at 22.5° intervals, thereby accounting for orientation variability [47]. As a result of these combined augmentation strategies, the expanded reference dataset used to train the kNN classifier increased to a total population of 22,865 images.

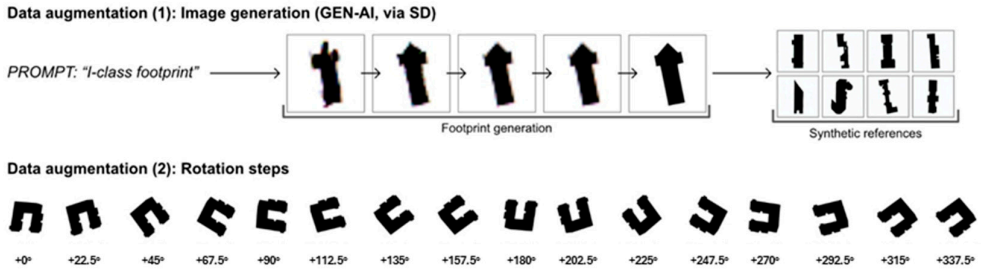


Figure 10. Data augmentation process.

For the kNN algorithm, Euclidean distance was chosen as the similarity metric, and distance-based weighting was applied to give greater importance to closer neighbors. This approach enhances the model's discriminative power by reducing the influence of more distant neighbors, which are less likely to reflect the local decision boundary accurately. The parameter "*k*" was initially set to 7, corresponding to the number of image classes to be recognized, but was later reduced to 5 after accuracy testing demonstrated improved overall performance with this setting. Lower values were not considered to avoid the risk of overfitting.

Figure 11 displays the confusion matrices corresponding to the selected parameters with $k = 5$ and $k = 7$.

k = 5

		Predicted							
		C	E	H	I	L	O	R	Σ
Actual	C	90.5 %	0.5 %	1.4 %	0.2 %	6.2 %	1.2 %	0.0 %	4.650
	E	8.8 %	81.1 %	6.7 %	0.2 %	1.8 %	1.4 %	0.0 %	1.215
	H	1.7 %	0.4 %	96.1 %	0.0 %	1.5 %	0.3 %	0.2 %	3.165
	I	0.4 %	0.1 %	0.2 %	95.4 %	1.5 %	0.0 %	2.4 %	3.135
	L	6.9 %	0.4 %	1.4 %	2.0 %	88.0 %	0.1 %	1.2 %	4.395
	O	4.2 %	0.2 %	0.6 %	0.0 %	0.4 %	94.6 %	0.0 %	3.705
	R	0.0 %	0.0 %	0.1 %	4.6 %	0.7 %	0.0 %	94.6 %	2.610
	Σ	4.839	1.046	3.278	3.216	4.305	3.589	2.602	22.875

k = 7

		Predicted							
		C	E	H	I	L	O	R	Σ
Actual	C	89.5 %	0.3 %	1.8 %	0.3 %	6.9 %	1.1 %	0.1 %	4.650
	E	11.1 %	75.1 %	9.6 %	0.2 %	2.3 %	1.6 %	0.0 %	1.215
	H	2.5 %	0.3 %	94.6 %	0.1 %	1.8 %	0.5 %	0.2 %	3.165
	I	0.3 %	0.1 %	0.4 %	95.0 %	1.9 %	0.0 %	2.4 %	3.135
	L	7.1 %	0.4 %	1.9 %	2.4 %	86.6 %	0.1 %	1.5 %	4.395
	O	5.0 %	0.3 %	0.8 %	0.0 %	0.5 %	93.3 %	0.0 %	3.705
	R	0.0 %	0.0 %	0.1 %	4.7 %	0.8 %	0.1 %	94.3 %	2.610
	Σ	4.886	969	3.322	3.225	4.311	3.550	2.612	22.875

Figure 11. Comparison of the confusion matrices for $k = 5$ and $k = 7$.

Each row of the matrices corresponds to the true class of the test images, while each column corresponds to the predicted class. Reading across a row shows how the images of a specific class are distributed across the predictions. The percentages in each row sum up to 100%, representing the total number of test images belonging to that class from the manual labeling process. However, the vertical sums across columns may exceed 100% because misclassified images from other classes can contribute to the same predicted category. The diagonal entries (highlighted in blue) indicate correct classifications, while off-diagonal entries (highlighted in red) represent misclassifications. The matrices were generated using a 5-fold cross-validation approach. In this method, the dataset is divided into five equally sized folds. At each iteration, four folds are used for training, while the remaining fold serves as the testing set. This process is repeated five times, with each fold serving as the validation set exactly once. The final accuracy is calculated as the sum of the classifications across all iterations, providing a more reliable and stable estimate of model performance compared to a single train–test split. Similar applications of cross-validation are commonly performed with 5 to 10 folds; however, in this case, the training dataset was relatively small, so 5 folds were selected to ensure enough samples in both the training and testing sets [48]. The confusion matrices map the proportion of actual samples that were correctly or incorrectly classified across all classes. With k = 5, an average accuracy of 91.4% was achieved, meaning that most images were correctly associated with their respective class. The reported average accuracy value was obtained by calculating the mean of the percentages of correctly classified images across all classes. This represents an improvement over the initial setting of k = 7, which produced an overall accuracy of 89.7% and exhibited a notably lower performance of 75.1% for class E. In addition, a performance comparison was conducted to evaluate the accuracy improvements provided by data augmentation. Figure 12 shows the resulting confusion matrices, comparing the performance of the kNN classifier trained on the manually labeled dataset (n = 1225) with that trained on the augmented dataset, expanded through both SD synthesis and rotation permutations (n = 22,875). The results highlight a substantial increase in classification performance: the average proportion of correctly classified images rose from 69.2% with the original dataset to 91.4% after data augmentation, demonstrating the effectiveness of this approach in enhancing model accuracy.

k = 5, without data augmentation (n = 1225)

		Predicted							
		C	E	H	I	L	O	R	Σ
Actual	C	65.3 %	0.0 %	5.0 %	1.8 %	18.7 %	8.7 %	0.5 %	219
	E	14.9 %	45.9 %	25.7 %	0.0 %	9.5 %	4.1 %	0.0 %	74
	H	11.0 %	1.2 %	71.2 %	1.8 %	9.8 %	3.7 %	1.2 %	163
	I	1.6 %	0.5 %	2.1 %	82.9 %	8.6 %	0.0 %	4.3 %	187
	L	18.4 %	1.3 %	3.3 %	8.4 %	66.1 %	0.8 %	1.7 %	239
	O	26.0 %	0.0 %	4.5 %	0.0 %	2.8 %	63.8 %	2.8 %	177
	R	1.2 %	0.0 %	1.2 %	6.0 %	1.8 %	0.0 %	89.8 %	166
	Σ	267	40	168	192	246	143	169	1.225

k = 5, with data augmentation (n = 22875)

		Predicted							
		C	E	H	I	L	O	R	Σ
Actual	C	90.5 %	0.5 %	1.4 %	0.2 %	6.2 %	1.2 %	0.0 %	4.650
	E	8.8 %	81.1 %	6.7 %	0.2 %	1.8 %	1.4 %	0.0 %	1.215
	H	1.7 %	0.4 %	96.1 %	0.0 %	1.5 %	0.3 %	0.2 %	3.165
	I	0.4 %	0.1 %	0.2 %	95.4 %	1.5 %	0.0 %	2.4 %	3.135
	L	6.9 %	0.4 %	1.4 %	2.0 %	88.0 %	0.1 %	1.2 %	4.395
	O	4.2 %	0.2 %	0.6 %	0.0 %	0.4 %	94.6 %	0.0 %	3.705
	R	0.0 %	0.0 %	0.1 %	4.6 %	0.7 %	0.0 %	94.6 %	2.610
	Σ	4.839	1.046	3.278	3.216	4.305	3.589	2.602	22.875

Figure 12. Comparison of confusion matrices for kNN classification using the two training datasets: above, manually labeled images only; below, enhanced through data augmentation to increase the number of samples.

The kNN classification was applied to the full set of retrieved building footprints (n = 4151) using the parameters (k = 5, Euclidean distance, distance-based weighting). For each classified image, the algorithm computes a class probability, representing the proportion of the k-nearest neighbors that belong to the predicted class. Across the dataset, the average value for the highest-class probability per footprint was 76%, indicating that, in most cases, a strong majority of the nearest neighbors aligned with the predicted label.

3.3. Final Dataset Creation and Relevant Features

To create a new dataset of geocoded and classified schools, the obtained building shapes were associated with the following:

- The footprint area extracted with QuickOSM, the number of floors, and the gross volume from the Open Data platform;
- The heating degree days and the climatic areas (D.P.R. 412/1993 [12]) of the related locations;
- The school type macro-category (kindergarten, primary, middle, or high school).

This final dataset enables the classification of the Italian school building stock according to type, shape, and school grade. It allows for a consideration of both the varying needs of students across age groups and the different strategies applicable in diverse climatic conditions.

From this analysis, it was possible to reconstruct a reference educational building (REB) for primary schools, which emerged as the most common school type in Italy, following the requirements of DM 75, since most of the buildings were constructed after the promulgation of this law, as is visible in Figure 1. This standard provides, for each school grade, the minimum requirements for the different functional uses.

4. Results and Discussion

This section illustrates the main results of the morphological classification, focusing on the distribution of school building shapes and their geometric features. It also presents the definition of a reference primary school model and discusses the implications for energy-retrofit strategies.

4.1. School Shape and Related Geometrical Features Definition

In Italy, it is common for a school structure to host two or more school institutions (e.g., a primary and a middle school in the same building), especially in smaller municipalities. Since the present study compares different databases and aims to establish a unique key to correlate each building with the specific school activities taking place, only buildings housing exactly one institution were selected. In practical terms, only the buildings where the ratio between the school code (*codicescuola*) and the building code (*codiceedificio*) is 1:1 were considered. After the geocoding process, it is not possible to determine whether a school building hosts one or more school institutions; therefore, this criterion enables a stronger and more certain association between the geocoded footprints database and the one containing the geometrical data, allowing the creation of more realistic and reliable school archetypes. Furthermore, this restriction ensured a clearer attribution of shape and layout characteristics to a specific school type, avoiding overlaps or hybrid configurations that may result from shared-use facilities.

Based on this filtered dataset, a typological matrix (Figure 13) was developed to represent the most common building shapes across school levels and Italian climatic areas.

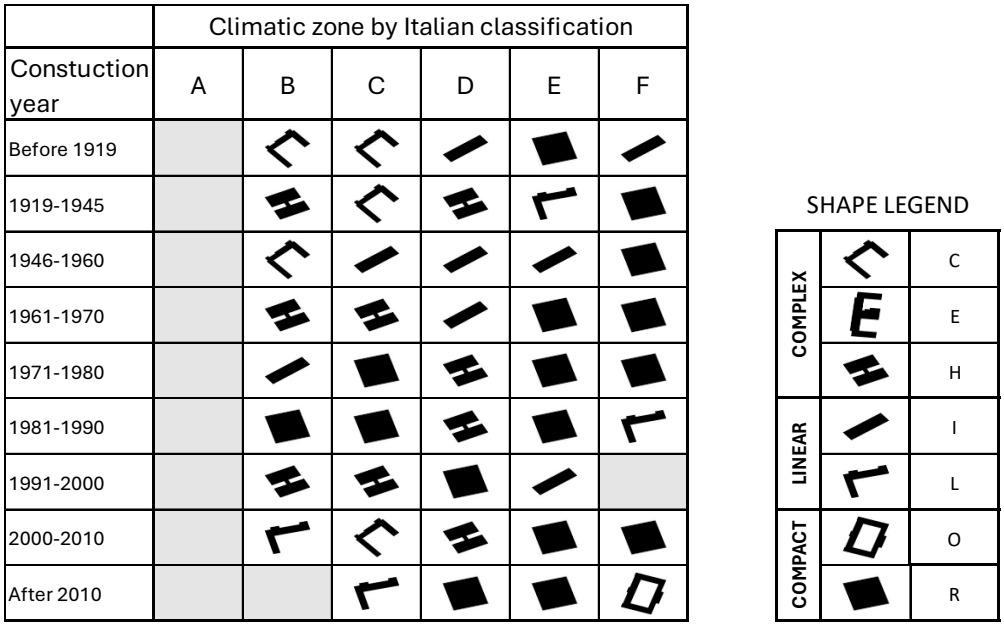


Figure 13. Typological matrix of primary school buildings, showing the most common shape by climatic area and construction year.

Coupled with the typological matrix in Figure 13, Table 3 shows the number of buildings in each category and key geometrical data (the median footprint area and the median number of floors) associated with each identified shape. Median values were used to provide a more representative overview of the dataset, as mean values can be skewed by outliers, whereas the median more reliably reflects the overall trend of the distribution. In addition, this approach provides contextual representativeness, which is essential for reconstructing representative models of the population.

Table 3. Geometric features of primary school by shapes.

Shape	Nr. of Buildings	Floors (Median)	Area (m^2—Median)
C	117	3	1542.3
E	18	3	1574.4
H	205	2	1708.9
I	227	3	897.7
L	223	3	1186.1
O	13	2	1565.3
R	319	2	713.6

In addition to this summary table, Appendix B provides a detailed statistical analysis of the number of floors, gross volume, and gross area, displaying violin plots with the most meaningful statistical indicators.

The classification process led to the identification of four recurrent building shapes for primary schools (H, I, L, and R) that can be further grouped into three sub-categories, based on similarities in spatial layout and expected thermal behavior:

- Linear shapes: I, L;
- Compact shapes: R, O;
- Complex shapes: H, C, E.

From this clustering, it emerged that primary schools in Italy have between two or three floors and a footprint area ranging from 714 m^2 to 1708 m^2, depending on the number of classrooms that a school can host.

4.2. Example of a Reference Primary School

The adopted approach led to a data-driven, structured definition of REBs based on actual building stock characteristics, providing a reliable foundation for further energy analysis, modeling, and intervention planning.

The school building graphically depicted in Figure 14 represents an example of a typical primary school compliant with DM 75 and characterized by an I-shape, as defined in the methodology section. According to this standard, primary schools are characterized by a minimum of 5 classrooms and a maximum of 25, corresponding to a minimum number of 75 students and a maximum of 625. In addition to this information, DM 75 specifies the key functional area requirements per destination of use (classroom, circulation zones, service, administrative offices, common areas like the canteen), stipulating a minimum area requirement per student of 5.21 m^2, with 1.8 m^2/student dedicated to classroom spaces. A detailed breakdown of surface requirements per student is provided in Table A2 of Appendix C, alongside Table A3, which lists the minimum internal height for the different functional areas.

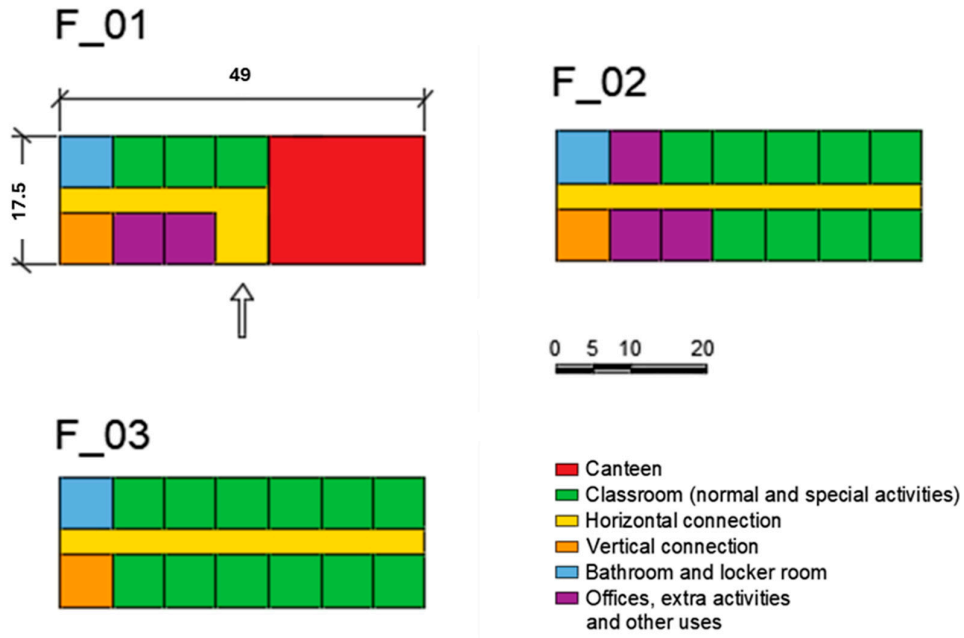


Figure 14. Example of a reference educational building for primary schools—I-shape. S/V: 0.36. Classrooms: 20. Expected students: 500.

The reconstruction is based on elementary modular units that reflect the spatial and functional organization suggested in the Italian decree.

The building consists of three floors above ground. The ground floor has a rectangular footprint measuring 49 × 17.5 m, yielding a total footprint area of 857.5 m^2.

Each floor plan is organized in a highly rational and functional manner, prioritizing accessibility, clarity of circulation, and modularity. The spaces are distributed according to function:

- Classrooms (green) are the predominant space and are distributed evenly across all levels;
- Horizontal and vertical circulation areas (yellow/orange) facilitate efficient circulation and accessibility, ensuring compliance with current standards;
- Administrative and complementary spaces (purple) include offices, support services, and areas for extracurricular activities;
- Restrooms and locker rooms (blue) are strategically positioned for shared access;

- The canteen (red) is situated on the ground floor for direct access from the main entrance.

The building's layout aligns with both pedagogical needs and regulatory compliance, offering a balance between learning environments, communal areas, and support spaces. Based on the total gross surface area of approximately 2572.5 m^2 across all three floors, the functional breakdown by area type is reported in Table 4.

Table 4. Internal space distribution of the reference educational building for the ground floor.

Function	Color	Area (m^2)	Percentage
Classrooms	Green	401.6	46.8%
Circulation (horizontal/vertical)	Yellow/Orange	253.5	29.6%
Restrooms and Locker Rooms	Blue		
Canteen	Red	115.2	13.4%
Administrative/Extra functions	Purple	87.2	10.2%
Total	–	857.5	100%

4.3. External Validity Considerations

In recent decades, building stock modeling has been widely used to quantify and evaluate the current and future energy and indoor environmental quality performance of large numbers of buildings at the city, regional, or national level. Building stock models commonly use building archetypes, which aim to represent the diversity of building stocks through frequently occurring building typologies.

The taxonomy proposed in this work could be slightly adapted to reflect international differences.

Adaptation to other countries would, at a minimum, requires the following:

- Mapping the buildings into appropriate climate bins (e.g., ASHRAE or EN ISO classifications) to account for boundary conditions that strongly influence indoor environmental quality and energy performance;
- Aligning the taxonomy with national or regional regulations and school building standards, which often determine performance thresholds for IEQ parameters.

A necessary starting point for any adaptation is the availability of open data on the school building stock. Several national databases already exist, covering aspects such as school registry information, geolocation, demographic data, and patterns of space use (e.g., in France [49], Germany [50], Spain [51], the Netherlands [52], Sweden [53], Australia [54], and New Zealand [55]). However, information on the construction characteristics and the conservation status of buildings is often more limited. More comprehensive datasets can be found in the United Kingdom [56] and in some states of Canada [57,58] and the U.S.A. [59], although not always in open-access format. In other contexts, a possible strategy would be to cross-reference geolocated school registry data with national building registries that cover the entire stock of real estate, including school facilities (e.g., Denmark [60]).

For adaptation to other public building typologies, the taxonomy should be revised by consulting state-of-the-art databases and literature for the target typology (e.g., hospitals, offices, libraries), to integrate use profiles, occupancy patterns, and specific regulatory requirements. Nevertheless, the hierarchical organization and logic of the taxonomy can be maintained, allowing comparability across building types.

5. Conclusions

This study presents a novel and replicable methodology for the large-scale morphological classification of buildings, applied in this case to the Italian school building stock, an area that could greatly benefit from the strategic planning of renovation interventions. A

clearer and more precise understanding of the morphological characteristics of large public building stocks can substantially enhance the accuracy of energy performance analyses, given the well-established influence of building shape on energy behavior.

Starting from an initial public dataset comprising over 40,000 school buildings, the research developed a taxonomy of recurring morphological types, offering a systematic and data-driven insight into the shape patterns of the existing educational stock. By employing k-nearest neighbors (kNN) image classification techniques, based on a combination of real, manually labeled school footprints and synthetically generated ones, typical Italian school shapes were successfully identified. Additionally, a reference educational building model for primary schools, the most widespread typology in Italy, was created according to reference regulatory standards to support energy performance simulations and assessments.

From a practical perspective, this taxonomy provides policymakers, designers, and public administrations with a strategic framework to prioritize and tailor retrofit interventions according to the specific morphological features of each building category. In doing so, it supports the development of more effective, scalable, and context-sensitive policies for the renovation and decarbonization of the educational sector at the national level.

This study relies on publicly available datasets, which may be incomplete, outdated, or lacking in critical information. Since the analysis is specifically tailored to the Italian school building stock, the findings may not be directly generalizable to other building types or national contexts. While morphological classification provides a valuable framework for organizing and understanding the building stock, it also introduces a certain level of simplification. In particular, the reliance on 2D building footprints, whether derived from GIS data or synthetically generated images, overlooks the vertical dimension of buildings (such as height and number of floors), which plays a crucial role in determining energy performance. This information would be derived from additional public datasets, although these sources may still be subject to the same limitations noted previously.

The clustering of data and image recognition using the k-nearest neighbors (kNN) algorithm is suitable for the specific task addressed in this study, given the task involved and the use of black-and-white images. Nevertheless, this method may be insufficient for more complex classification tasks or higher-dimensional datasets, where advanced deep learning models would likely offer better performance.

A further limitation lies in the absence of a direct connection between the defined morphological categories and actual measured energy performance data. This gap limits the ability to validate the real impact of retrofit strategies on energy efficiency. Although the study includes the development of a reference model for primary schools (Italy's most prevalent typology), it may not fully capture the diversity of local architectural variants. However, the intention was to provide an average and representative contribution, leveraging a broader dataset than those typically used in the literature, while recognizing the inherent challenge of obtaining detailed information for such a large number of schools.

Future research could focus on the systematic identification of reference energy buildings (REBs) representative of the wide variety of school buildings, considering differences in school grades, construction typologies, and climatic zones. This would enable the creation of benchmark models that reflect the real diversity of the educational building stock. In parallel, it would be essential to identify typical renovation interventions associated with each REB cluster, thus supporting the development of targeted, scalable, and cost-effective deep renovation strategies aligned with national and European decarbonization goals.

As future developments of this research will include not only energy performance but also comfort analyses, the adoption of parametric studies will be required to systematically explore multiple scenarios. Consequently, parameters such as orientation, boundary conditions, and possible alternative functions were intentionally left undefined at this stage. In

line with current literature emphasizing the decisive role of building form and the surface-to-volume ratio (S/V), the definition of solid and reliable three-dimensional archetypes constitutes a significant outcome of the present work. These archetypes provide a robust and reliable foundation for subsequent investigations, enabling both energy and comfort evaluations across a wide range of parametric configurations and operational conditions.

Author Contributions: Conceptualization, G.C., M.C., F.R.C., A.G.M., M.M.S. and E.D.G.; methodology, G.C., M.C., F.R.C., A.G.M., M.M.S. and E.D.G., software; G.C., M.C. and F.R.C., validation, G.C., M.C., F.R.C. and A.G.M.; formal analysis, G.C., M.C., F.R.C. and A.G.M.; investigation, G.C., M.C., F.R.C. and A.G.M.; resources, F.R.C., A.G.M., M.M.S. and E.D.G.; data curation, G.C., M.C., F.R.C. and A.G.M.; writing—original draft preparation, G.C., M.C., F.R.C. and A.G.M.; writing—review and editing, G.C., M.C., F.R.C., A.G.M., M.M.S. and E.D.G.; visualization, G.C., M.C. and F.R.C.; supervision, F.R.C., A.G.M., M.M.S. and E.D.G.; project administration, A.G.M., M.M.S. and E.D.G.; funding acquisition, A.G.M., M.M.S. and E.D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the Ministero dell'Università e della Ricerca (MUR) by PRIN: Progetti di Rilevante Interesse Nazionale—Call 2022—Prot. 20228RX7R8, CUP:I53C24002540006.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GIS	Geographic Information System
kNN	k-nearest neighbors
HVAC	Heating, ventilation, and air conditioning
EPBD	Energy Performance of Buildings Directive
HDD	Heating degree day
GG	*Gradi Giorno*
S/V	Surface-to-volume ratio
EUI	Energy use intensity
RSE	*Ricerca Sistema Energetico*
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
MBF	Microsoft Building Footprint
OSM	OpenStreetMap
SD	Stable diffusion
DM 75	*Decreto Ministeriale 18 dicembre 1975, n. 29*

Appendix A

Every municipality in Italy is classified into one of six climatic zones (A to F) based on heating degree days, a parameter that quantifies the heating demand of a location. This classification, established by D.P.R. 412/1993 [12] and referenced in the "Decreto Requisiti Minimi" [61], plays a fundamental role in national energy regulations, as different thermal transmittance limits (U-values) for building envelope components are prescribed for each climatic zone, ensuring that energy efficiency measures are appropriate to the local climate. In addition, the classification is used to regulate the permissible operation periods of heating systems, with maximum daily operating hours and activation periods defined according to the zone's severity.

Climatic zones are determined by the annual sum of the positive differences between a base temperature (20 °C) and the daily mean outdoor temperature, on all days when the average temperature falls below this threshold, expressed in degree days (Table A1).

Table A1. Italian climatic zones' classification.

Climatic Zone	Heating Degree Days (HDDs)
A	<600
B	$600 \leq$ HHD < 900
C	$900 \leq$ HHD < 1400
D	$1400 \leq$ HHD < 2100
E	$2100 \leq$ HHD < 3000
F	≥3000

The Italian school system is divided into four main educational stages, each tailored to specific age groups and learning goals, listed as follows.

Kindergarten: attended by children aged 3 to 5, this stage is not mandatory but widely attended.

Primary school: starting at the age of 6 and lasting until the age of 10, this five-year compulsory stage introduces students to the fundamentals of reading, writing, mathematics, science, history, and geography.

Middle school: for students aged 11 to 13, it is compulsory and lasts three years. It offers more structured, subject-specific instruction.

High school: designed for students aged 14 to 18, this five-year stage offers different pathways, as there is the possibility to choose between academic (e.g., classic, scientific, artistic schools), technical, and professional schools.

Appendix B

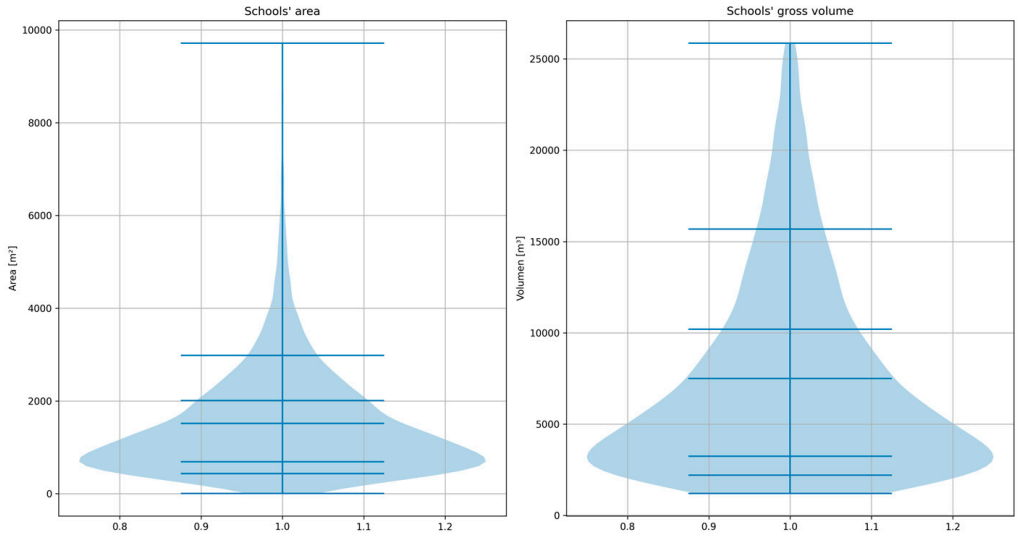


Figure A1. Gross dimensions of the classified primary schools.

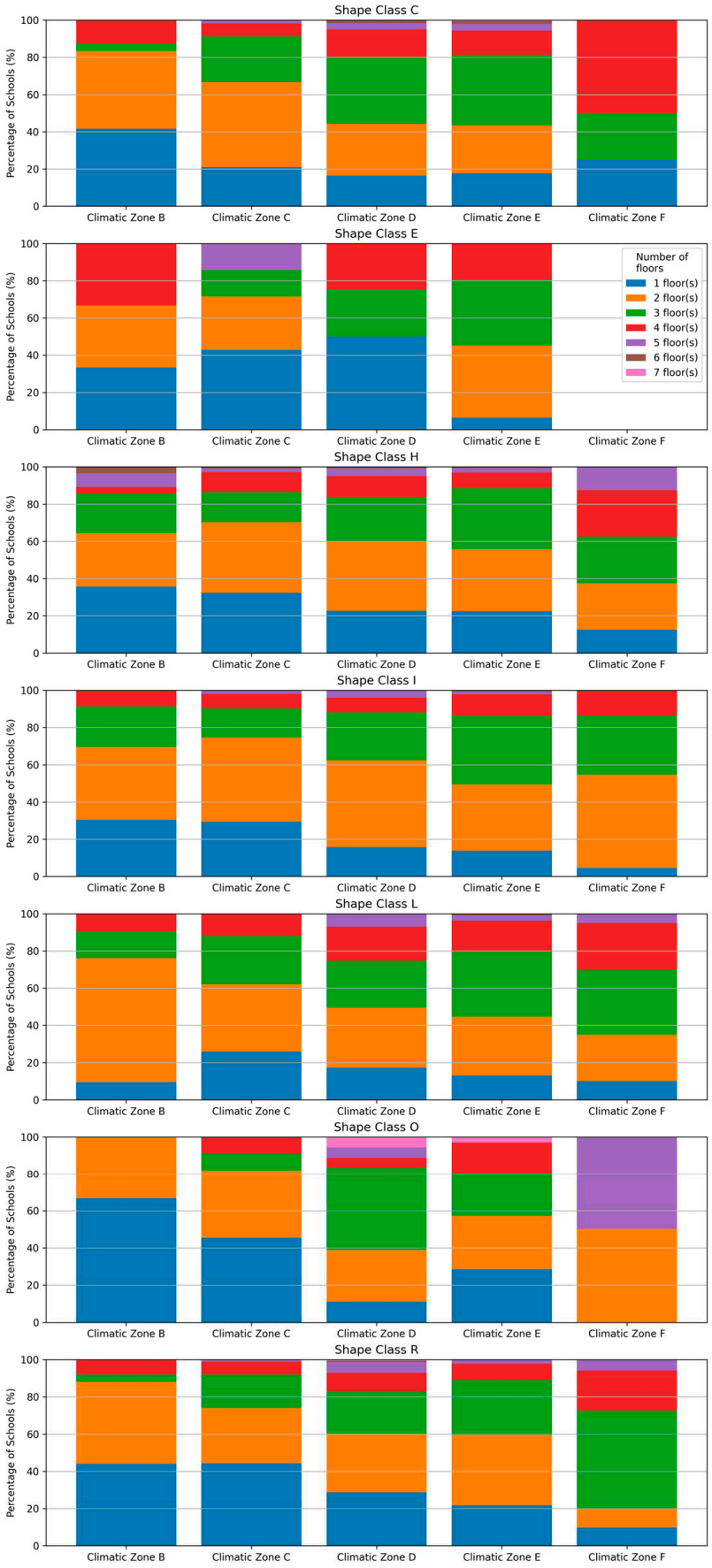


Figure A2. Detailed breakdown of the number of floors by climatic zone and building shape.

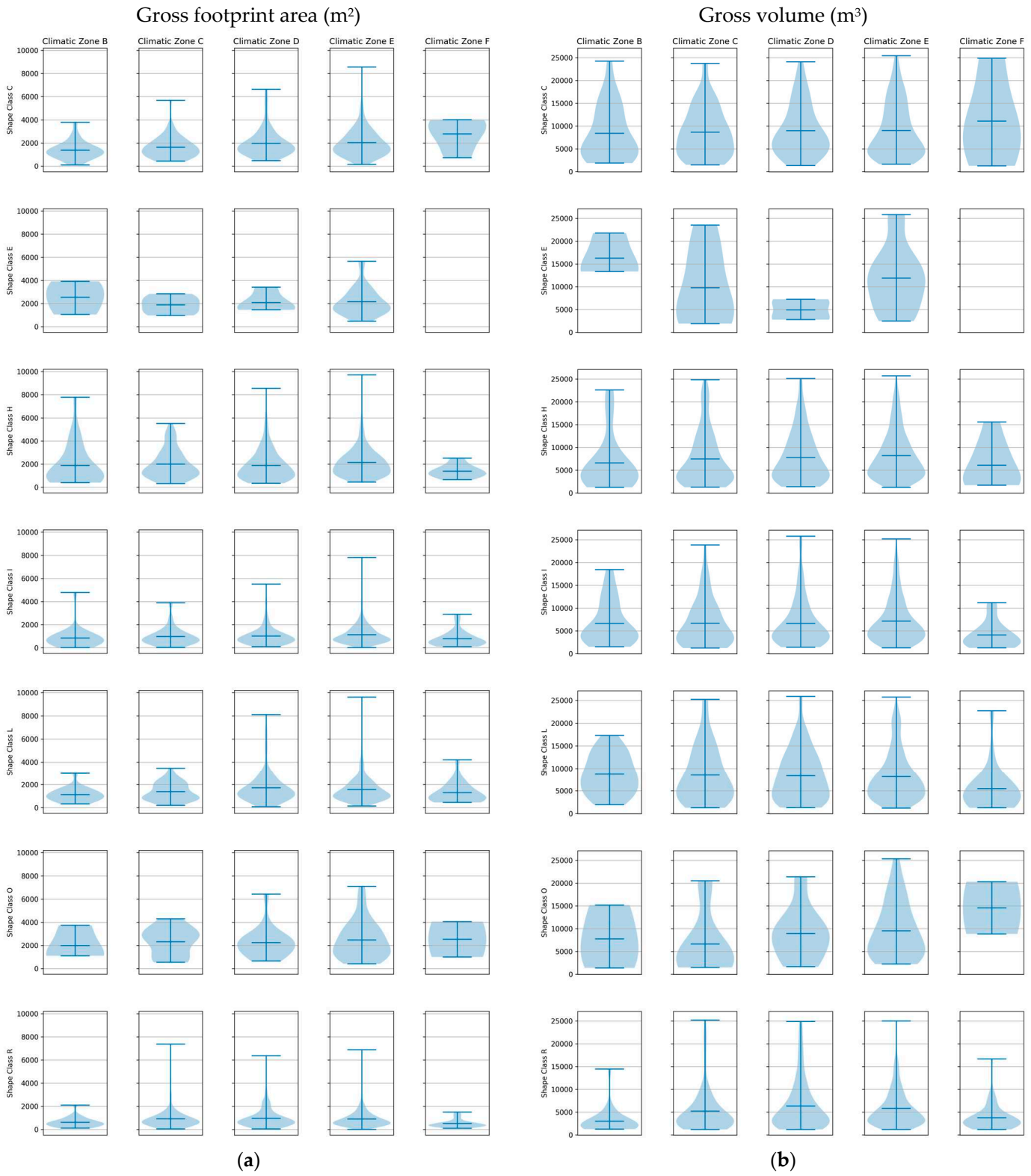


Figure A3. Detailed gross footprint area (**a**) and gross volume (**b**) by climatic zone and building shape.

Appendix C

Table A2. DM 75 area requirements for primary school.

Primary School	
Activity Description	**m^2/Student**
Teaching activities	
Regular activities	1.80
Inter-cycle activities	0.64
Total surface index related to teaching activities	
Min.	2.44
Max	2.70
Collective Activities	
Supplementary and extracurricular activities:	0.40
Canteen and related services	0.70
Complementary Activities	
Teachers' library	0.13
Sum of partial indices	
Min.	3.67
Max.	3.93
Circulation and sanitary facilities (42% of the previous sum)	
Min.	1.54
Max.	1.65
Total net surface index	5.21
Maximum total net surface index	5.58

Table A3. DM 75 minimum floor height requirements.

	Type of Space	Minimum Height (cm)	Notes
1	Pedagogical unit spaces (classroom)	300	With flat ceiling. In case of sloped ceiling, minimum height 270 cm.
	Group work areas	240	
2	Specialized teaching spaces	300	With flat floor and ceiling.
	If tiered: lowest part	240	
3	Laboratories and workshops		According to specific regulations.
4	Communication and information spaces:		
	(i) Library	300	
	Carrel area	210	
	(ii) Auditorium and multipurpose activity hall:		
	If tiered: lowest part	240	
	Highest part	420	
	Without tiers	420	
5	Physical education spaces:		In case of an A2-type gymnasium with a volleyball court installation (point 3.5.1.), the minimum height must be 720 cm.
	Type A gymnasium	540	
	Type B gymnasium	720	
6	Circulation spaces	240	
7	Administrative and medical examination spaces	300	
8	Dining spaces:		
	(a) if in a niche up to 30/35 m^2	240	
	(b) in all other cases	300	

References

1. Premrov, M.; Žigart, M.; Leskovar, V.Ž. Influence of the building shape on the energy performance of timber-glass buildings located in warm climatic regions. *Energy* **2018**, *149*, 496–504. [CrossRef]
2. ReRoads Project. Available online: https://reroads-project.unibs.it/ (accessed on 31 July 2025).
3. Commission, E. Energy Performance of Buildings Directive (EBPD) IV. Available online: https://energy.ec.europa.eu/topics/energy-efficiency/energy-performance-buildings/energy-performance-buildings-directive_en (accessed on 31 July 2025).
4. Ciulla, G.; D'Amico, A. Building energy performance forecasting: A multiple linear regression approach. *Appl. Energy* **2019**, *253*, 113500. [CrossRef]
5. Chen, K.W.; Janssen, P.; Schlueter, A. Multi-objective optimisation of building form, envelope and cooling system for improved building energy performance. *Autom. Constr.* **2018**, *94*, 449–457. [CrossRef]
6. Parasonis, J.; Keizikas, A.; Kalibatiene, D. The relationship between the shape of a building and its energy performance. *Archit. Eng. Des. Manag.* **2012**, *8*, 246–256. [CrossRef]
7. del Merito, M.D.E. Portale Unico dei Dati della Scuola—Open Data—Edilizia Scolastica. Available online: https://dati.istruzione.it/opendata/opendata/catalogo/elements1/?area=Edilizia%20Scolastica (accessed on 29 June 2025).
8. Alghamdi, A.; Hu, G.; Haider, H.; Hewage, K.; Sadiq, R. Benchmarking of water, energy, and carbon flows in academic buildings: A fuzzy clustering approach. *Sustainability* **2020**, *12*, 4422. [CrossRef]
9. Salvalai, G.; Malighetti, L.E.; Luchini, L.; Girola, S. Analysis of different energy conservation strategies on existing school buildings in a Pre-Alpine Region. *Energy Build.* **2017**, *145*, 92–106. [CrossRef]
10. Marrone, P.; Gori, P.; Asdrubali, F.; Evangelisti, L.; Calcagnini, L.; Grazieschi, G. Energy benchmarking in educational buildings through cluster analysis of energy retrofitting. *Energies* **2018**, *11*, 649. [CrossRef]
11. Pedone, L.; Molaioni, F.; Vallati, A.; Pampanin, S. Energy refurbishment planning of Italian school buildings using data-driven predictive models. *Appl. Energy* **2023**, *350*, 121730. [CrossRef]
12. Decreto del Presidente della Repubblica 26 Agosto 1993, n. 412. 1993. Available online: https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg (accessed on 30 June 2025).
13. Decreto Ministeriale 18 Dicembre 1975. Available online: https://www.edscuola.it/archivio/norme/decreti/dm181275.html (accessed on 30 June 2025).
14. Marincu, C.; Dan, D.; Moga, L. Investigating the influence of building shape and insulation thickness on energy efficiency of buildings. *Energy Sustain. Dev.* **2024**, *79*, 101384. [CrossRef]
15. Baglivo, C.; Albanese, P.M.; Congedo, P.M. Relationship between shape and energy performance of buildings under long-term climate change. *J. Build. Eng.* **2024**, *84*, 108544. [CrossRef]
16. Torabi, M.; Simonen, K.; Evins, R. What matters the most in designing low-carbon buildings in Canada? Exploring the tradeoff between embodied and operational carbon in early stage design. *Energy Build.* **2025**, *334*, 115482. [CrossRef]
17. Kistelegdi, I.; Horváth, K.R.; Storcz, T.; Ercsey, Z. Building Geometry as a Variable in Energy, Comfort, and Environmental Design Optimization—A Review from the Perspective of Architects. *Buildings* **2022**, *12*, 69. [CrossRef]
18. Li, J.; Liang, C.; Zhou, W. A Review of Building Physical Shapes on Heating and Cooling Energy Consumption. *Energies* **2024**, *17*, 5766. [CrossRef]
19. PRISMA Statement. Available online: https://www.prisma-statement.org/ (accessed on 1 July 2025).
20. Energetico, R.S.S. I Consumi Della Pubblica Amministrazione. 2024. Available online: https://www.rse-web.it/wp-content/uploads/2025/01/2025-RSEview-PA.pdf (accessed on 12 September 2025).
21. Campagna, L.M.; Fiorito, F. On the energy performance of the Mediterranean school building stock: The case of the Apulia Region. *Energy Build.* **2023**, *293*, 113187. [CrossRef]
22. Kazem, H.A.J.; Al-Kazzaz, D.A. Modelling Design Standards for Iraqi Schools Using Building Information Modeling. *Int. J. Sustain. Dev. Plan.* **2023**, *18*, 1477–1487. [CrossRef]
23. Bo, C.; De Angelis, E.; The Management of the Energy Performance Simulation of a Complex Building Portfolio. The Case of the School Building Asset of an Italian Municipality. 2022. Available online: https://pro.unibz.it/library/bupress/publications/fulltext/9788860461919_19.pdf (accessed on 11 September 2025).
24. Geraldi, M.S.; Gnecco, V.M.; Neto, A.B.; Martins, B.A.d.M.; Ghisi, E.; Fossati, M.; Melo, A.P.; Lamberts, R. Evaluating the impact of the shape of school reference buildings on bottom-up energy benchmarking. *J. Build. Eng.* **2021**, *43*, 103142. [CrossRef]
25. Zinzi, M.; Pagliaro, F.; Agnoli, S.; Bisegna, F.; Iatauro, D. On the built-environment quality in nearly zero-energy renovated schools: Assessment and impact of passive strategies. *Energies* **2021**, *14*, 2799. [CrossRef]
26. Ignjatovic, N.D.C.; Ignjatovic, D.M.; Zekovic, B.D. Improving energy efficiency of kindergartens in Serbia: Challenges and potentials. *Therm. Sci.* **2020**, *24*, 3521–3532. [CrossRef]
27. Akil, M.; Tittelein, P.; Defer, D.; Suard, F. Statistical indicator for the detection of anomalies in gas, electricity and water consumption: Application of smart monitoring for educational buildings. *Energy Build.* **2019**, *199*, 512–522. [CrossRef]

28. Zhang, A.; Bokel, R.; van den Dobbelsteen, A.; Sun, Y.; Huang, Q.; Zhang, Q. The effect of geometry parameters on energy and thermal performance of school buildings in cold climates of China. *Sustainability* **2017**, *9*, 1708. [CrossRef]
29. Hussein, M.H.; Barlet, A.; Baba, M.; Semidor, C. Evaluation for Environmental Comfort Performance in the Palestinian Schools. In Proceedings of the PLEA 2016 Los Angeles—36th International Conference on Passive and Low Energy Architecture-Cities, Buildings, People: Towards Regenerative Environments, Los Angeles, CA, USA, 11–13 July 2016; Available online: https://www.researchgate.net/publication/308762434 (accessed on 11 September 2025).
30. Lara, R.A.; Pernigotto, G.; Cappelletti, F.; Romagnoni, P.; Gasparella, A. Energy audit of schools by means of cluster analysis. *Energy Build.* **2015**, *95*, 160–171. [CrossRef]
31. Morck, O.; Romeo, C.; Zinzi, M. On the implementation of an innovative energy/financial optimization tool and its application for technology screening within the EU-project School of the Future. In *Energy Procedia*; Elsevier Ltd.: Amsterdam, The Netherlands, 2015; pp. 3330–3335. [CrossRef]
32. Santamouris, M.; Mihalakakou, G.; Patargias, P.; Gaitani, N.; Sfakianaki, K.; Papaglastra, M.; Pavlou, C.; Doukas, P.; Primikiri, E.; Geros, V.; et al. Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy Build.* **2007**, *39*, 45–51. [CrossRef]
33. Ivanova, E. School building planning. Main types of systems (plans) of school buildings. *World Sci.* **2019**, *1*, 18–31. [CrossRef]
34. Rigolon, A. European Design Types for 21st Century Schools. 2010. Available online: https://www.oecd.org/en/publications/european-design-types-for-21st-century-schools_5kmh36gpvmbx-en.html (accessed on 11 September 2025).
35. Sole, M. *Manuale di Edilizia Scolastica*; Carocci: Roma, Italy, 1995.
36. di Bitonto, A.; Giordano, F. *L'architettura Degli Edifici per L'istruzione*; Officina Edizioni: Rome, Italy, 1995.
37. Dahlström, L.; Johari, F.; Broström, T.; Widén, J. Identification of representative building archetypes: A novel approach using multi-parameter cluster analysis applied to the Swedish residential building stock. *Energy Build.* **2024**, *303*, 113823. [CrossRef]
38. Lucchi, E.; D'Alonzo, V.; Exner, D.; Zambelli, P.; Garegnani, G. A density-based spatial cluster analysis supporting the building stock analysis in historical towns. In Proceedings of the Building Simulation Conference Proceedings, International Building Performance Simulation Association, Rome, Italy, 2–4 September 2019; pp. 3831–3838. [CrossRef]
39. QuickOSM. Available online: https://plugins.qgis.org/plugins/QuickOSM/ (accessed on 1 July 2025).
40. Zhou, Y.; Leung, Y.; Zhang, W.-B. A Location-and-Form-Based Distance for Geographical Analysis. *Ann. Am. Assoc. Geogr.* **2021**, *111*, 1253–1270. [CrossRef]
41. Ujianto, N.T.; Gunawan; Fadillah, H.; Fanti, A.P.; Saputra, A.D.; Ramadhan, I.G. Penerapan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi citra medis. *IT-Explor. J. Penerapan Teknol. Inf. Dan Komun.* **2025**, *4*, 33–43. [CrossRef]
42. Jia, M.; Chen, B.-C.; Wu, Z.; Cardie, C.; Belongie, S.; Lim, S.-N. Rethinking Nearest Neighbors for Visual Classification. *arXiv* **2021**, arXiv:2112.08459.
43. Abubakar, F.A.; Boukari, S. A Convolutional Neural Network with K-Neareast Neighbor for Image Classification. *Int. J. Adv. Res. Comput. Commun. Eng.* **2018**, *7*, 1–7. [CrossRef]
44. Bhat, A.D.; Acharya, H.R.; Srikanth, H.R. A Novel Solution to the Curse of Dimensionality in Using KNNs for Image Classification. In Proceedings of the 2019 2nd International Conference on Intelligent Autonomous Systems, ICoIAS 2019, Singapore, 28 February–2 March 2019; Institute of Electrical and Electronics Engineers Inc.: New York City, NY, USA, 2019; pp. 32–36. [CrossRef]
45. Nijhawan, R.; Verma, M.; Miglani, M.K. Satellite Image Classification Through Stable Diffusion and Vision Transformers. In Proceedings of the 2025 3rd International Conference on Disruptive Technologies, Greater Noida, India, 7–8 March 2025; Institute of Electrical and Electronics Engineers Inc.: New York City, NY, USA, 2025; pp. 871–875. [CrossRef]
46. Nerogar/OneTrainer. Available online: https://github.com/Nerogar/OneTrainer (accessed on 4 September 2025).
47. Perez, L.; Wang, J. The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. *arXiv* **2017**, arXiv:1712.04621.
48. Aprihartha, M.A.; Idham, I. Optimization of Classification Algorithms Performance with k-Fold Cross Validation. *Eig. Math. J.* **2024**, *7*, 61–66. [CrossRef]
49. Plateforme ouverte des données Éducation, Sports et Jeunesse. Available online: https://data.education.gouv.fr/pages/accueil/ (accessed on 28 August 2025).
50. Kultusministerkonferenz—Schulstatistik. Available online: https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik.html (accessed on 28 August 2025).
51. la Transformación Digital y de la Función Pública, M.P. datos.gob.es. Available online: https://datos.gob.es/es/sector/educacion (accessed on 28 August 2025).
52. DUO Open Onderwijsdata. Available online: https://duo.nl/open_onderwijsdata/ (accessed on 28 August 2025).
53. Statistik om Förskola Och Skola. Available online: https://www.skolverket.se/statistik-och-utvarderingar/statistik-om-forskola-och-skola/sok-statistik-om-forskola-skola-och-vuxenutbildning?sok=SokC (accessed on 28 August 2025).
54. ACARA-Australian Curriculum Assessment and Reporting Authority. My School. Available online: https://www.myschool.edu.au/ (accessed on 28 August 2025).

55. Education Counts. Available online: https://www.educationcounts.govt.nz/home (accessed on 28 August 2025).
56. Condition Data Collection 2 (CDC2) Programme. Available online: https://www.gov.uk/guidance/condition-data-collection-2-cdc2-programme (accessed on 28 August 2025).
57. School Facility Condition Information (SFCI). Available online: https://data.ontario.ca/en/dataset/school-facility-condition-information-sfci (accessed on 28 August 2025).
58. School Facility Inventory System (SFIS). Available online: https://data.ontario.ca/en/dataset/school-facility-inventory-system-sfis (accessed on 28 August 2025).
59. Washington Office of Superintendent of Public Instruction. Information and Condition of Schools (ICOS). Available online: https://ospi.k12.wa.us/policy-funding/school-buildings-facilities/information-and-condition-schools-icos (accessed on 28 August 2025).
60. Sonderborg Kommune. Available online: https://sonderborgkommune.dk/ (accessed on 28 August 2025).
61. Decreto Interministeriale 26 Giugno 2015. Available online: https://www.mimit.gov.it/it/normativa/decreti-interministeriali/decreto-interministeriale-26-giugno-2015-applicazione-delle-metodologie-di-calcolo-delle-prestazioni-energetiche-e-definizione-delle-prescrizioni-e-dei-requisiti-minimi-degli-edifici (accessed on 18 July 2025).