

Towards Automated Occupant Profile Creation in Smart Buildings: A machine learning-enabled approach for user persona generation

Sheik Murad Hassan Anik*, Xinghua Gao[†] and Na Meng*

*Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

[†]Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA 24061, USA

Abstract—The user persona is a communication tool for designers to generate a mental model that describes the archetype of users. Developing building occupant personas is proven to be an effective method for human-centered smart building design, which considers occupant comfort, behavior, and energy consumption. Optimization of building energy consumption also requires a deep understanding of occupants' preferences and behaviors. The current approaches to developing building occupant personas face a major obstruction of manual data processing and analysis. This study proposes a machine learning-based approach for occupant characteristics classification and prediction with a view toward partially automating the building occupant persona generation process. It investigates the 2015 Residential Energy Consumption Dataset with six machine learning techniques — Linear Discriminant Analysis, K Nearest Neighbors, Decision Tree (Random Forest), Support Vector Machine, and, AdaBoost classifier — for the prediction of 16 occupant characteristics, such as age, education, and, thermal comfort. The models achieved moderate accuracy in predicting most of the occupant characteristics and significantly higher accuracy (over 90%) for attributes including the number of occupants in the household, their age group, and preferred usage of primary cooling equipment. The results of the study show the feasibility of using machine learning techniques for occupant characteristics prediction and automating the development of building occupant persona to minimize human effort.

I. INTRODUCTION

People spend most of the time indoors [1], and the impact of a building on its occupants is significant [2]. Smart Buildings are buildings that integrate intelligence, enterprise, control, materials, and construction as an entire building system, with adaptability, not reactivity, at its core, to meet the drivers for building progression: energy and efficiency, longevity, comfort, and, satisfaction [3]. In recent years, human-centeredness has become an increasingly important factor for smart building design and operation. Optimizing building performance requires a deep understanding of occupants' behavior and preferences. Developing building occupant personas is proven to be effective in creating occupant profiles for human-centered smart buildings [4]. An occupant persona is a tool of communication for designers to develop a conceptual model that describes the archetype of occupants [5]. The development of a building occupant persona can assist the building designers, developers, and, engineers in accurately estimating resource demand through realistic and representative occupant profiles, enabling pragmatic occupant-centric building design

and operation. The increasing availability of data from a wide range of sources will allow smart buildings to become adaptable, and prepare themselves for context and change over all timescales [3].

One of the key objectives of smart buildings is to maximize occupant comfort while minimizing energy consumption [6]. Buildings and urban spaces increasingly incorporate artificial intelligence and new forms of interactivity, raising a wide span of research questions about the future of human experiences with, and within, built environments. This emerging area is defined as Human-Building Interaction (HBI) [7]. HBI affects human well-being and the surrounding environment. Although energy and technology are placed into buildings by people for their own comfort and recreation, often the decision-making process excludes the human factor from the design itself. Smart buildings provide an opportunity to consider the physical, physiological, and, psychological needs of occupants in the building design. Human-centered design (HCD) places humans at the center of the building design, which can accelerate smart housing design for people. To maximize human well-being and the optimize performance of smart buildings, an iterative, human-centered approach to building design must be employed.

Agee *et al.* [4] introduced a human-centered approach to smart housing, leading to the development of data-driven smart housing personas that communicate smart housing user needs. In their work, the authors utilized descriptive statistics and behavioral analysis to describe the physical, physiological, and psychological needs of the occupants. Personas help designers anchor their work in a fictional user's needs in the design of products and systems [8]. Their proposed approach can reduce the risk of designers designing for themselves.

Developing building occupant personas benefits both the occupants and building designers in many ways. A user-centric building design can help understand the needs, behaviors, and preferences of the potential occupants, building designers can create spaces that are tailored to the users' needs, leading to a more user-centered design approach. By designing buildings with the occupant persona in mind, the building can provide a more pleasant and satisfying experience for the occupants,

leading to higher levels of user satisfaction and productivity.

Designing buildings that meet the needs of the occupant persona can lead to improved building performance, such as energy efficiency and indoor air quality, as the building is designed to meet the specific needs of its occupants. It can also reduce the need for retrofits and renovations, resulting in cost savings over the building's lifecycle. Understanding the needs and preferences of potential occupants can help building-owners and developers to better market and lease the building, leading to higher occupancy rates and increased revenue. However, the task faces multiple challenges beginning with the limited availability of building occupant data. The lack of data related to occupants and their living environment is one of the major pullbacks in smart housing persona development research. Even with the limited data that is available, the process of developing building occupant persona is manual and time-consuming. Currently, the occupant persona development process requires researchers manually analyzing data, conducting interviews, clustering the occupant profiles based on multiple criteria, and finally constructing a persona [4]. The entire process takes much time because it involves manual labor in all the steps. Automating a single step of this procedure can accelerate the entire persona development task. Moreover, although there is a scarcity of residential building data, there has been some work to address this issue recently. For example, Anik *et al.* [9] presented a cost-effective and portable framework for indoor data collection, and Song, *et al.* [10] discussed data collection and analysis methods for analyzing human comfort in the indoor environment. With more building data available, and the development of machine learning technologies, new opportunities are emerging. Machine learning is an automated process that extracts patterns from data [11]. In the field of predictive data analytics, machine learning is a method used to devise complex prediction algorithms and models [12]. Machine learning models that can deliver fast and accurate results given appropriate data are provided [13].

In this work, we apply machine learning to the 2015 Residential Energy Consumption Survey Data for building occupant characteristics classification and prediction which can lead to automating some steps of building occupant persona development. The procedure includes data processing, feature engineering, selection of target variables, training, and evaluating the models. This work aims to answer the following research questions:

- RQ1: How effectively can machine learning tools predict individual building occupant characteristics?
- RQ2: How do machine learning algorithms compare with each other when predicting building occupant characteristics?

The rest of the paper is structured as follows: Section II reviews relevant studies conducted in the domain of building occupant persona development and machine learning methods

for occupant behavior modeling. Section III presents the methodology, description of the data, and machine learning models used. Section IV demonstrates the machine learning evaluation details. Section V discusses the development of building occupant persona using the findings of this work, the limitations, and, the future research directions. Finally, section VI concludes the research.

II. RELATED WORKS

Related previous studies can be grouped into the following two categories: Building energy performance with respect to occupant behavior and Machine learning in occupant behavior modeling.

A. Building energy performance and occupant behavior

HBI usually reflects some characteristics of the building occupants, and the building occupant persona is an effective and concise way to represent these characteristics for human-centered building design and operation. Several studies have investigated the relationship between occupant behavior and building performance, focusing on different aspects and contexts.

Agee *et al.* [4] and Malik *et al.* [14] examined the occupants' behaviors and characteristics in residential housing. Agee *et al.* conducted a study on a human-centered approach to smart housing, collecting data from 309 residential housing units in Virginia, U.S. They employed a multi-phase, mixed-methods research design and used affinity diagramming to identify occupants' HBI activities and develop data-driven Personas. Malik *et al.* researched the occupant behavior of 1223 low-income households in India and established occupant archetypes based on behavioral and psychographic characteristics. Both studies emphasize the importance of understanding occupant behavior for energy demand estimation and occupant-centric building designs.

O'Brien *et al.* [15] and D'Oca *et al.* [16] focused on integrating occupant dimensions into building design and performance assessment. O'Brien *et al.* proposed occupant-centric building performance metrics, while D'Oca *et al.* reviewed the "human dimensions" of energy use in buildings and advocate for their integration into design and operation processes. These studies highlighted the need to consider occupant perspectives and behaviors in building performance evaluation and energy reduction initiatives.

Pan *et al.* [17] and Sun *et al.* [18, 19] explored occupant behavior and its impact on energy efficiency. Pan *et al.* analyzed occupant behavior and electricity load patterns in Shanghai, China, identifying different groups with distinct patterns and recommending energy-saving measures. Sun *et al.* addressed the challenge of accurately evaluating the energy savings potential of occupant behavior and propose frameworks for quantifying the influence of occupant behavior on energy conservation measures. Both studies provided valuable insights for energy efficiency decision-making and building energy simulation.

Dong *et al.* [20, 21] focused on occupant behavior modeling and control strategies. Dong *et al.* presented a methodology for integrated building heating and cooling control using predictive models of occupant behavior, while their second work explores the potential of utilizing urban big data and advanced modeling methods for modeling urban-scale occupant behaviors. These studies contributed to the development of occupant behavior models and control strategies for reducing energy consumption while maintaining occupant comfort.

Hu *et al.* [22] emphasized the importance of integrating occupant behavior into building energy policies, discussing the challenges and opportunities in this area. They highlighted the need for further research and consider the integration of occupant behavior in building technical standards, regulations, and energy incentives. The study provided insights into the role of occupant behavior in energy policy development.

Ortiz *et al.* [23] studied the motivations behind comfort behaviors and energy consumption discrepancies among occupants with different behavioral patterns. They grouped occupants into categories based on psychological and behavioral models, highlighting the distinct opinions and discrepancies in self-reported answers. The study emphasized the importance of understanding occupants' comfort needs for a better understanding of home energy consumption.

Buttitta *et al.* [24] proposed a method for developing occupancy-integrated archetypes to characterize heat demand in residential buildings. They show that using occupancy profiles from existing models may lead to significant discrepancies in heat demand estimation. The study highlights the need for differentiated and disaggregated energy profiles in building energy modeling.

B. Machine learning in occupant behavior modeling

Amasyali *et al.* [25] and Carlucci *et al.* [26] both employed machine learning approaches to model occupant behavior and its influence on energy consumption. Amasyali *et al.* proposed a data-driven method that includes machine learning-based models for predicting energy consumption and comfort, along with a genetic algorithm-based optimization model for optimizing occupant behavior. Carlucci *et al.* used a machine learning model trained on a large dataset of energy-use cases to predict building energy consumption, highlighting the impact of occupant behavior and identifying opportunities for energy-saving measures. Li *et al.* [27] utilized machine learning techniques to develop a load prediction model for residential buildings, considering occupant behavior as a predictor variable. Their study demonstrated the potential of data-driven machine learning models in aiding decision-making for building design and retrofit processes.

Deng *et al.* [28] focus on predicting indoor thermal comfort by developing artificial neural network (ANN) models that incorporate thermal sensations and occupants' behavior. The models were trained on data collected from offices and houses/apartments, showing promise in predicting thermal comfort in different environments. Peng *et al.* [29] proposed

a demand-driven control strategy for HVAC systems that responds to occupants' energy-related behavior, aiming to reduce energy consumption and maintain room temperature. Their approach combined unsupervised and supervised machine learning methods to learn occupants' behavior and automate the cooling system's control. Kim *et al.* [30] explored the development of personal comfort models based on occupants' heating and cooling behavior data. They employed machine learning algorithms and demonstrated improved accuracy in predicting individuals' thermal preferences compared to conventional models. The work showed the potential of the models to enhance occupant satisfaction and energy efficiency in buildings.

Sun *et al.* [31] introduced a data analytics framework for detecting changes in occupant behaviors to support the design of energy-saving incentives. Their framework combined forecasting individual energy consumption with determining behavior change probabilities, utilizing machine learning and statistical methods. The study emphasized the importance of advanced analytics for detecting behavior changes and designing effective energy-saving measures. Yu *et al.* [32] presented a methodology for identifying and improving occupant behavior in residential buildings. Their approach incorporated cluster analysis, classification analysis, and association rules mining to analyze end-use loads and identify energy-inefficient behaviors. The methodology proved effective in providing recommendations to occupants and evaluating the energy-saving potential of behavior modifications.

These studies collectively contribute to understanding occupant behavior, its influence on building performance, energy consumption, comfort, and, the potential of machine learning and data analytics in modeling and optimizing occupant behavior for energy-efficient buildings. The development of a human-centered building and understanding of building occupants requires vast data on both energy consumption and occupant behavior. The absence of either of these factors not only obstructs the thoroughness of the process of developing a persona for building occupants but also leads to inaccurate outcomes. Despite the challenges in modeling occupant behavior due to the unpredictable nature and diversity of human behavior, it is crucial to examine the general patterns of their behaviors and incorporate this information into the building energy model [33]. Modeling occupant behavior remains one of the key steps in the process of building occupant persona development. A valid occupant behavior model needs to have the potential to simulate realistic building users' reactions to different built environments, and the lack of real occupant behavioral data accounts for a gap in research. The human-centered building occupant persona development process proposed in [4] requires manual human work of labeling, filtering, and clustering data from large datasets. These steps are both labor and time extensive, which impede the entire process. To address these research gaps, this research utilizes real-world occupant data with machine learning tools to classify building

occupant characteristics which can accelerate the process of building occupant persona development.

III. METHODOLOGY

Developing occupant characteristics persona is a difficult and time-consuming task [4]. Using the latest machine learning tools can help alleviate the process. However, there are some challenges to using machine learning tools in the development of building occupant persona. This section describes the challenges and the methodology followed by this work to overcome these challenges in using state-of-the-art machine learning models in occupant characteristics classification and prediction.

A. Challenges of using machine learning tools in building occupant persona development process

This work aimed towards using state-of-the-art machine learning models in the development of building occupant personas. Along the path, the hurdles in the task were identified. The following are the key challenges of using machine learning models in building occupant persona development:

(i) Acquiring the necessary data: Obtaining a dataset that accurately represents the target population can be challenging and time-consuming. Collecting data on a variety of factors such as demographics, lifestyle, and behavior can be costly and may require specialized knowledge.

(ii) Identifying relevant variables: Determining which variables to include in the model can be difficult due to the large number of potential variables that may be relevant. This requires a deep understanding of the target population and the factors that influence their behavior.

(iii) Choosing the appropriate model: Selecting which ML models to use: There are many different machine learning models that can be used for occupant persona development, each with its own strengths and weaknesses. Selecting the most appropriate model for a given problem requires a thorough understanding of the data and the problem at hand. Additionally, different models may require different levels of computational resources and expertise to implement, which can affect the feasibility of using a particular model.

(iv) Fine-tuning hyperparameters of models for individual target variables: Finally, even after selecting an appropriate machine learning model, there is a significant amount of tuning that must be done to ensure that the model is effective at predicting the target variables. This involves adjusting the hyper-parameters of the model to optimize performance, which can be time-consuming and require significant computational resources. Moreover, hyperparameter tuning may need to be done separately for each target variable, which further increases the complexity of the problem.

(v) Interpreting the results: Interpreting the results of a machine learning model can be challenging, especially when dealing with complex models. Understanding the factors that contribute to the model's predictions can be difficult, which may limit the usefulness of the model for developing occupant personas.

(vi) Data privacy concerns: Collecting and storing data on individuals raises privacy concerns. Care must be taken to ensure that data is collected and used in compliance with relevant privacy laws and regulations.

B. Methodology Overview

Figure 1 illustrates the framework of this research, which is inspired by a study conducted by Zhongguo *et al.* [34]. The process begins with processing the 2015 Residential Energy Consumption Dataset [35]. The metadata is extracted, and the data is cleaned in the pre-processing step. Then, 16 target variables are chosen to represent occupant characteristics, such as age, education, income, thermal preference, etc. The rest of the attributes remain as descriptive variables and are filtered through the feature selection step, in which the irrelevant and redundant variables are dropped. Then, six machine learning models for classification prediction tasks are selected and tested. These models are trained through the 10-fold cross-validation and the evaluated results are presented here. The following sections describe the different components of the methodology of this work.

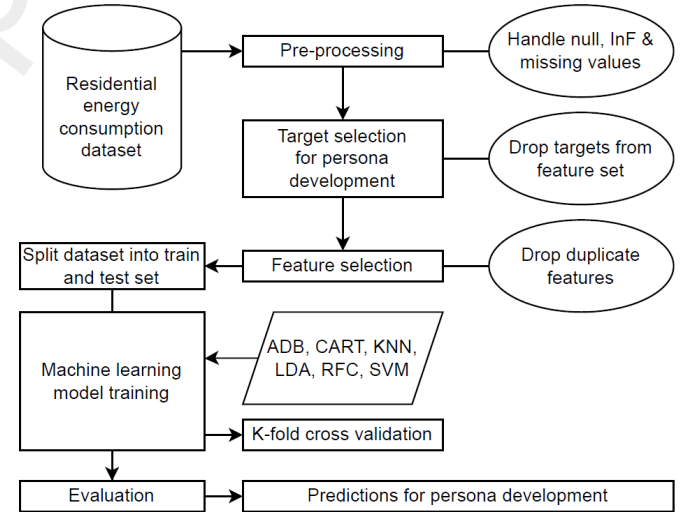


Fig. 1. Our research methodology involving data curation, ML model training, and ML-based prediction for persona development

C. Dataset description

The U.S. Energy Information Administration (EIA) conducts the Residential Energy Consumption Survey (RECS) periodically to gather detailed information on energy usage in American households [35]. RECS is a comprehensive project that includes a household survey, data

collection from energy suppliers, and estimation of end-use consumption and expenditures. The survey gathers data on energy-related features and usage patterns from a nationally representative sample of housing units. The energy supplier survey collects data on the consumption of electricity, natural gas, propane/LPG, fuel oil, and kerosene in the sampled housing units during the reference year. RECS focuses on homes used as primary residences and excludes secondary homes, vacant units, military barracks, and common areas in apartment buildings. Consequently, RECS estimates are not meant to represent total sector-level figures but are particularly useful for comparing different characteristics of homes within the residential sector.

The total number of responding households is 5686 in the 2015 RECS. Each record includes a total of 759 attributes. These attributes have been categorized in 12 sections by EIA:

- A: Structural Characteristics – house type, construction time, renovation time, floor plan, building materials, etc.
- B: Kitchen Appliances – kitchen stove, microwave, stove fuels, oven, hood, ventilation, grill, refrigerator, freezer, dishwasher, etc.
- C: Home Appliances and Electronics – clothes washer, dryer, washing cycle, television, gaming console, phone, computer, etc.
- D: Space Heating – heating appliances, heater usage during winter, heating fuel, thermostat controls, hot water system, fireplace, etc.
- E: Air Conditioning – air conditioners, cooling systems, air conditioning heat pump, air filter, programmable thermostat, thermostat control during summer days, etc.
- F: Water Heating – water heaters types, fuels, usage, age, tank, etc.
- G: Miscellaneous – light bulbs, swimming pool, energy audit, energy assistance, etc.
- H: Fuels Used – back-up generator, onsite power system details, fuel usage in home, energy bills, natural gas usage in home, etc.
- I: Housing Unit Measurement – size, shape and area of the different floors of the housing units.
- J: Fuel Bills – fuel suppliers, bill types, different fuel bills like electricity, gas, wood, smart meter, etc.
- K: Housing Unit Characteristics – occupants' age, gender, education, employment status, number of people living in the household, income, etc.
- L: Energy Insecurity and Assistance – challenges paying bills, struggles due to unsafe or unhealthy temperature, utility discontinuation, energy assistance, requirement of medication, etc.

Figure 2 illustrates the number of features in each of the 12 categories in pairs. For every pair, the left bar (colored in blue) refers to the original number of features in that category, and, the right bar (colored in orange) refers to the number of features selected after the feature selection process. 16 of these attributes are chosen to be the target variables because they

relate to individual occupant characteristics. The rest of the attributes remain as input variables. The next section discusses the feature selection process in detail. In the training process, the k-fold cross-validation is utilized with $k = 10$ dividing the 5686 records into 10 random folds where 9 folds (5118 data points) were used for training and the remaining 1 fold of data (568 data points) was used for testing. The process is repeated 10 times to ensure the validity of the evaluation results.

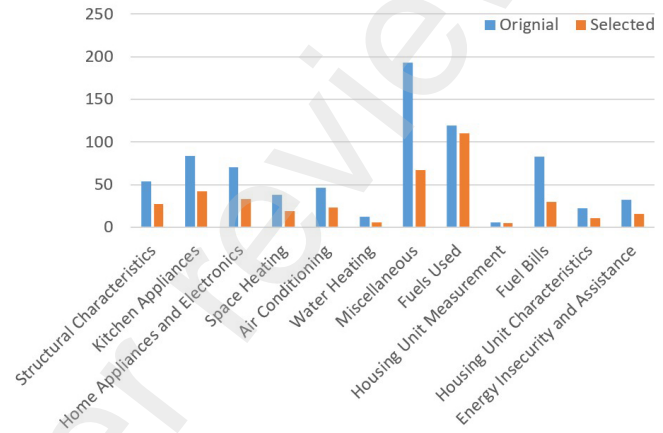


Fig. 2. Feature distribution in categories.

D. Pre-processing

Data pre-processing is the process of transforming raw data into a machine-understandable format. It is a critical step in any machine-learning pipeline as it can have a significant impact on the performance of the model. In many cases, the raw data used for machine learning may be incomplete, noisy, or contain errors. Sometimes, such noisy data make the model unusable. Pre-processing can help to clean and correct these issues, which can improve the accuracy of the model. This work utilizes the data from RECS 2015 dataset where most of the columns are ready to be directly fed to the machine learning models. However, there are some columns that contain infinite, missing, blank, and, null values. For machine learning models to understand all records, these values needed to be transformed. The following are the different pre-processing steps done in this work:

1) *Missing, null, InF values:* Some of the data records contained missing or blank values. Some records had null values. These missing, blank, and, null values have been replaced with -1 to establish the value is missing. Some rows included infinite values written as 'inf' instead of a number. These records are particularly difficult to process by the machine learning models because the model expects all values of a particular column to be of the same data type. Having both numbers and text or string on the same column provides a mixed signal to the machine learning pipeline. The infinite values have been replaced with a large enough number, i.e. the largest integer in the programming language Python.

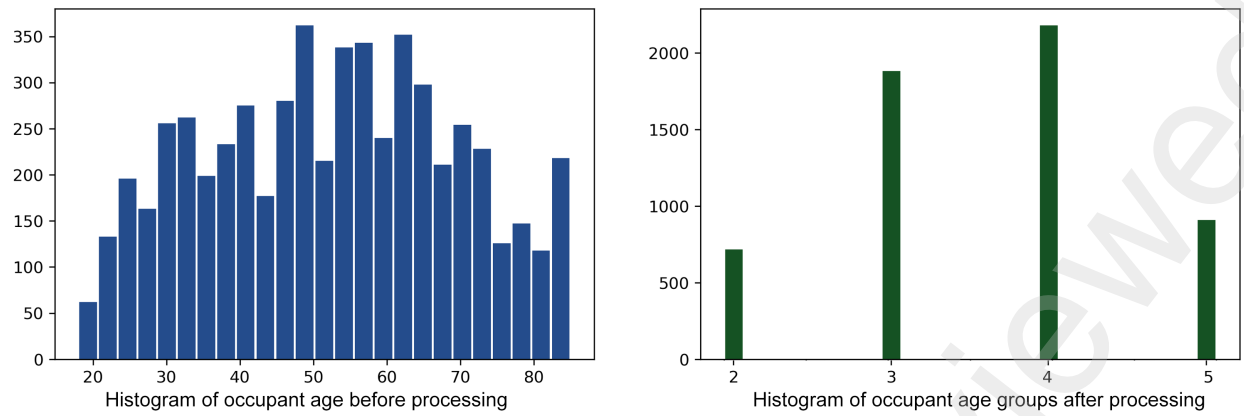


Fig. 3. Histogram of occupant age data before (left) and after (right) pre-processing step.

2) *Age classification*: The target variable, HHAGE ((SL. 9 on Table I)) which refers to the age of the responding occupant, is continuous from 18 to 110. Previous studies [36, 37] used age groups instead of continuous age numbers in years which is more feasible in machine learning classification tasks. This study follows the age range used in [37]. The records are categorized in the following age groups (in years): Children (0 to 12), Young Adult (13 to 30), Middle Adult (31 to 50), Senior Adult (51 to 70), and Senior (71 to 110). Figure 3 illustrates the histogram of the occupant age data before (left) and after (right) the pre-processing step. On the left, the data is distributed across 18 to 110 years. On the right, there are only 4 bars representing each age group. As the occupant age data starts from 18 years, there are no records in the Children age group (0 to 12 years).

3) *Temperature classification*: The temperature attributes (SL. 3 to 8 on Table I) are in a continuous form representing temperature records in degree Fahrenheit (°F). Similar to the age group, it is preferable to use regression models for continuous numbers. In the case of a classification task, it is better to have groups or classes of numbers. In the RECS 2015 dataset, the temperature data in all 6 target columns range from 40 to 96 °F. As an example of the data distribution of the temperature records, the left side of Figure 4 illustrates the histogram of the attribute TEMPGONE which represent the temperature data when no one is at home during the winter days. The other temperature attributes demonstrate similar characteristics. Here, it can be observed that most of the data fall between 50 to 80 °F. However, there is a peak on 0 °F. This portion represents missing or null data. The continuous temperature data is then distributed in 7 groups. Details of these groups can be found on SL. 3 of Table I. All temperature attributes have been categorized following the same distribution. This categorization is more feasible for machine learning classification tasks. In this way, machine learning models are used to predict a temperature group for a particular input instead of identifying the exact temperature value. On the right side of Figure 4 the histogram after the processing of TEMPGONE attribute is shown. Here, it can

be seen that most of the data lie in groups 4 and 5 which represent 65 to 68°F and 69 to 72°F respectively.

E. Feature selection

Feature selection is an important step in the performance of effective machine learning models. It involves selecting a subset of the most relevant features (or variables) from a larger set of available features that are likely to have the most predictive power for the target variable. This process can lead to better model performance, faster training times, avoiding over-fitting, and, more interpretable models.

The RECS 2015 dataset includes a total of 759 columns each representing an attribute. The authors went through a manual inspection of each of the features to identify the features that are important in the occupant characteristics prediction and the features that can be removed to reduce the total number of features to feed the machine learning models. A total of 370 attributes were dropped through the manual feature selection process leaving 389 attributes for the target and training features.

The RECS 2015 dataset included imputation flags for most of the attributes which refer to whether a record was imputed or not. A record can be imputed if it is not directly obtained from the occupants, but it is observed from other sources such as the structure, surrounding, or measuring device. Either way, in this work, the imputation flag is not a necessary component. Hence, the imputation flags are dropped. There were a total of 217 imputation flag attributes.

The energy consumption attributes recorded in the dataset contained both energy units and corresponding costs in US dollars. The dollar amounts are not in the scope of this study and thus are dropped from the dataset. There were a total of 53 dollar amount attributes. The data also included replicated weights which are used for variance estimation. These attributes were also not in the scope of this study and thus a total of 97 replicated weight attributes were dropped

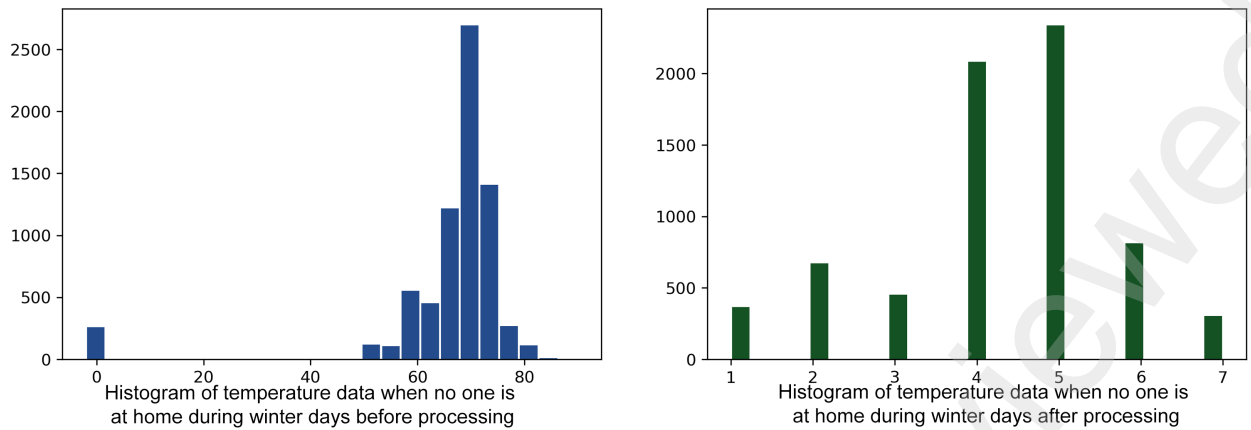


Fig. 4. Histogram of TEMPGONE before (left) and after (right) pre-processing step.

from the dataset.

Each record in the dataset was given an ID represented in the first column named DOEID. The ID is unique to each record and unnecessary in the classification task. It was dropped in the selection process. Finally, there were 2 attributes representing the number of phones and cellphones. These 2 columns have a direct correlation with the number of occupants living in the household which is one of the characteristics this work aims to estimate. Thus these 2 attributes were also dropped from the dataset.

Each of the 759 attributes were manually examined in this study and a total of 370 features were dropped in the manual feature selection process which includes imputation flags, utility bills in US dollar amount (actual unit is kept), and replicated weights (used for variance estimation), ID, and, phone counts. These are safe to drop because they do not represent any data that might affect the output of the machine-learning models for developing building occupant personas. The orange bars in Figure 2 (right bar of each category pair) represent the number of features selected for the training models. The remaining 389 attributes include features from all the categories of the initial dataset. After cautious examination of these 389 features, 16 target features were selected. More details on the target features are provided in the next section. Leaving the target features aside, there remains a total of 373 attributes which were kept as the input variables of the machine-learning models. This study utilized all of the remaining 373 features as inputs of the machine-learning models.

F. Target variables

The authors examined the remaining 389 attributes after the feature selection step to select the target variables. A total of 16 attributes were found that actively relate to occupant characteristics. These attributes can provide key information in developing building occupant persona. Table I provides a detailed overview of these 16 target attributes. The first

column (SL.) is a regular serial number of these variables, the second column (Attribute Name) denotes the respective target variable, the third column (Description) provides a description of the attribute, the fourth column (Data Range) represents the range span of the data of the respective attribute, the fifth column (# of classes) denotes the total number of classes present in that attribute, and, the last column (Class Details) provides detail information of the classes present in the particular target attribute. The following are the 16 target variables selected for this study:

- EQUIPMUSE: Main heating equipment household behavior, including values such as setting one temperature and leaving it there most of the time; manually adjusting the temperature at night or when no one is at home; programming the thermostat to automatically adjust the temperature during the day and night at certain times; turn equipment on or off as needed, etc. The datapoints are distributed in 6 classes. The class details can be found on the first row of table I.
- USEWWAC: Most-used individual air conditioning unit household behavior including values, such as setting one temperature and leaving it there most of the time; manually adjusting the temperature at night or when no one is at home; programming the thermostat to automatically adjust the temperature during the day and night at certain times; turn equipment on or off as needed, etc. Similar to EQUIPMUSE, the data points are categorized in 6 classes.
- TEMPHOME: Winter temperature when someone is at home during the day. The values range from 50 to 90 degrees Fahrenheit. The temperature attributes have been processed in the pre-processing step. The processed data falls into 7 classes. The class details can be found on row 3 of table I.
- TEMPGONE: Winter temperature when no one is at home during the day with the value range of 50 to 90 degrees Fahrenheit. It was also processed in the pre-processing step to distribute the data into 7 classes. Class

distribution is the same as TEMPHOME.

- TEMPNITE: Winter temperature at night with the value range of 50 to 90 degrees Fahrenheit. A similar processing step has been followed here with 7 classes.
- TEMPHOMEAC: Summer temperature when someone is at home during the day with the value range of 50 to 90 degrees Fahrenheit. The same temperature processing was conducted in this attribute and the data was distributed into 7 classes.
- TEMPGONEAC: Summer temperature when no one is at home during the day with the value range of 50 to 90 degrees Fahrenheit. Similar to the other temperature attributes, this one was also processed and the data is classified into 7 classes.
- TEMPNITEAC: Summer temperature at night with the value range of 50 to 90 degrees Fahrenheit. A similar processing step has been followed here with 7 classes.
- HHAGE: Respondent age, values ranging from 18 to 110. This attribute was processed in the pre-processing step and the data has been classified into 5 classes, from children to seniors. The class details are provided in table I.
- EMPLOYHH: Respondent employment status. Values covering employed full-time, part-time and unemployed or retired. Data is classified into 4 classes, starting from 1 to 4.
- EDUCATION: Highest education completed by the respondent. Replies cover less than high school diploma or GED, high school diploma or GED, some college or associate's degree, bachelor's degree (for example: BA, BS), master's, professional, or doctorate degree (for example: MA, MS, MBA, MD, JD, PhD). Data is distributed in 5 classes ranging from 1 to 5. Detailed class distribution can be found on row 11 of table I.
- NHSLDMEM: Number of household members, values ranging from 1 to 20. Each occupant number was automatically considered as a class by the machine learning models thus making the total number of classes to 20 for this attribute.
- NUMADULT: Number of household members age 18 or older, values ranging from 1 to 20. Again, each occupant number was automatically considered as a class by the machine learning models thus making the total number of classes 20 for this attribute.
- NUMCHILD: Number of household members age 17 or younger, values ranging from 1 to 20. Similar to the other attribute featuring occupant number, this attribute also contains 20 classes each representing a number between 1 to 20.
- ATHOME: Number of weekdays someone is at home. The values range from 0 to 5, creating a total of 6 classes.
- MONEYPY: Annual gross household income for the previous year. The data is distributed in 16 classes ranging from less than 5,000 USD to more than 150,000 USD. Detailed class distribution can be found on row 16 of table I.

These target attributes were chosen to describe different occupant characteristics. Together, these attributes can provide information on the age, lifestyle, behavior, family size, income, education, and, thermal comfortability of the occupants. The goal of this study is to see if we can predict these occupant characteristics from the selected 373 input features that contain respective information on energy usage, housing type, kitchen appliances, electronics usage, heating, and, cooling elements. If the models can successfully predict the target occupant characteristics from these selected input features, it can benefit in identifying the needs of particular residents from the data of their current household.

G. Machine learning models

Classification is a supervised learning approach in which a target variable is categorical or discrete. The task of choosing a specific classification model is a critical step [38], and each model has its own strengths and weaknesses in a given scenario. There is no cut-and-dried flowchart that can be used to determine which model should be used or will outperform the rest. A simple example can be, back propagation neural networks achieve higher accuracy than the decision tree method on Iris and Appendicitis data but a lower accuracy on Breast cancer and Thyroid data [34, 39]. Pedregosa *et al.* [40] suggests a guide on choosing the proper machine learning algorithm for a given task. In this work, six machine learning models have been used parallelly on the library recommended settings [40]. The following are the classification models used in this work:

1) *Linear Discriminant Analysis (LDA)*: Linear Discriminant Analysis (LDA) is a supervised learning algorithm used for classification tasks [40]. It is a technique used to find a linear combination of features that best separates two or more classes of objects or events. It can be used to classify data into different categories based on their features. It is a linear model for classification and is most commonly used for feature extraction in pattern classification problems. It is used in finding the projection hyper-plane that minimizes the inter-class variance and maximizes the distance between the projected means of the classes [41].

In LDA, it is assumed that the classes have the same covariance matrix and that the data is normally distributed. The goal is to find a projection of the data onto a lower-dimensional space, that is, a subspace such that the different classes are well separated and the intra-class variance is minimized. This projection is obtained by finding the eigenvectors of the between-class scatter matrix and the within-class scatter matrix. The between-class scatter matrix measures the variation between different classes, while the within-class scatter matrix measures the variation within each class. The eigenvectors of the between-class scatter matrix and the within-class scatter matrix are used to construct a transformation matrix that maps the data to a new subspace. In the new subspace, each sample is represented by a new

TABLE I
TARGET ATTRIBUTE DETAILS.

SL.	Attribute Name	Description	Data Range	# of classes	Class Details
1	EQUIPMUSE	Main heating equipment control description	1 to 5, & 9	6	1: Set one temperature and leave it there most of the time 2: Manually adjust the temperature at night or when no one is at home 3: Program the thermostat to automatically adjust the temperature during the day and night at certain times 4: Turn equipment on or off as needed 5: Our household does not have control over the equipment 9: Other
2	USEWWAC	Main cooling equipment control description	1 to 5, & 9	6	Same as 1 (EQUIPMUSE)
3	TEMPHOME	Typical temperature when someone is at home during the winter days	40 to 96 °F	7	1: 40 to 50 °F 2: 51 to 60 °F 3: 61 to 64 °F 4: 65 to 68 °F 5: 69 to 72 °F 6: 73 to 76 °F 7: 77 to 96 °F
4	TEMPGONE	Typical temperature when no one is at home during the winter days	40 to 96 °F	7	Same as 3
5	TEMPNITE	Typical temperature when someone is at home during the winter nights	40 to 96 °F	7	Same as 3 (TEMPHOME)
6	TEMPHOMEAC	Typical temperature when someone is at home during the summer days	40 to 96 °F	7	Same as 3 (TEMPHOME)
7	TEMPGONEAC	Typical temperature when no one is at home during the summer days	40 to 96 °F	7	Same as 3 (TEMPHOME)
8	TEMPNITEAC	Typical temperature when someone is at home during the summer nights	40 to 96 °F	7	Same as 3 (TEMPHOME)
9	HHAGE	Age of the respondent	18 to 110 years	5	1: Children (0 to 12) 2: Young Adult (13 to 30) 3: Middle Adult (31 to 50) 4: Senior Adult(51 to 70) 5: Senior (71 to 110)
10	EMPLOYHH	Employment Status of the respondent	1 to 4	4	1: Employed full-time 2: Employed part-time 3: Retired 4: Not employed
11	EDUCATION	Highest degree or level of school completed by the respondent	1 to 5	5	1: Less than high school diploma or GED 2: High school diploma or GED 3: Some college or Associate's degree 4: Bachelor's degree (e.g.: BA, BS) 5: Master's or higher degree (e.g.: MS, MBA, MD, PhD)
12	NHSLDMEM	Number of people living in the home	1 to 20	20	Each number of people is converted into a class itself
13	NUMADULT	Number of adults living in the home	1 to 20	20	Each number of people is converted into a class itself
14	NUMCHILD	Number of children living in the home	1 to 20	20	Each number of people is converted into a class itself
15	ATHOME	Number of weekdays someone is at home	0 to 5	6	0: None 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days
16	MONEYPY	Combined household income (USD)	1 to 16	16	1: Less than 5,000 2: 5,000 - 7,499 3: 7,500 - 9,999 4: 10,000 - 12,499 5: 12,500 - 14,999 6: 15,000 - 19,999 7: 20,000 - 24,999 8: 25,000 - 29,999 9: 30,000 - 34,999 10: 35,000 - 39,999 11: 40,000 - 49,999 12: 50,000 - 59,999 13: 60,000 - 74,999 14: 75,000 - 99,999 15: 100,000 - 149,999 16: 150,000 or more

set of features, which are called discriminant functions. These functions can be used to classify new data points by assigning them to the class with the highest score. This study used the LDA model with its default parameters [40]. The solver parameter was set to "svd" indicating Singular value decomposition (default). This is recommended for data with a large number of features which is the case of this study.

2) *K Nearest Neighbors Classifier (KNN)*: K Nearest Neighbors Classifier (KNN) is also a supervised machine learning algorithm used for classification tasks [40]. It is a non-parametric method, which means it does not make any

assumptions about the underlying distribution of the data. Instead, it uses the nearest neighbors of a new data point to predict its class. The k-NN algorithm is among the simplest and yet most efficient classification rules and is widely used in practice [42].

The KNN algorithm works by finding the K training examples that are closest to a new test example in terms of some distance metric, for example, Euclidean distance, and, cosine distance. The value of K is a hyperparameter that differs in different tasks. Once the K nearest neighbors are identified, the algorithm assigns the new test example to the

class that is most common among its K nearest neighbors. One of the strengths of KNN is that it can handle non-linear decision boundaries and can be applied to both binary and multi-class classification problems which is the case of this study. In this study, the parameter `n_neighbors` (also referred to as K) was set to 5. This is also the default recommended value of the parameter.

3) *Decision Tree Classifier (CART)*: Decision Tree Classifier (CART) is another supervised learning algorithm used for classification and regression tasks. It is a tree-based model that works by recursively partitioning the feature space into regions, each of which is associated with a class label or a continuous target value [43]. The CART algorithm builds a binary tree in which each internal node corresponds to a decision rule based on the value of a single feature. The decision rule splits the data into two or more subsets based on a threshold value or a set of categorical values. Each leaf node of the tree corresponds to a predicted class or target value. To build a decision tree, the CART algorithm uses a greedy approach that maximizes the information gain or the Gini impurity [44] at each split. Information gain measures the reduction in entropy (or some other measure of uncertainty) of the target variable, while Gini impurity measures the probability of misclassification if a random example is assigned to a particular class. In this work, the Gini impurity parameter was used as criteria for the split with the splitter strategy set to "best".

The CART algorithm can be used for both binary and multi-class classification problems, as well as for regression tasks. As the tasks related to this study are multi-class classification, CART remains a good fit. It is a simple and interpretable model that can handle both categorical and continuous features, and, can capture non-linear relationships between the features and the target variable.

4) *Support Vector Machine (SVM)*: Support Vector Machine (SVM) is a popular supervised learning algorithm used for classification, regression, and outlier detection tasks [45]. SVM aims to find the hyperplane that best separates the examples of different classes in a high-dimensional feature space. The basic idea of SVM is to map the input data points into a high-dimensional feature space using a kernel function, and then find the hyperplane that maximizes the margin between the two closest examples of different classes. The examples that are closest to the hyperplane are called support vectors, and they determine the position and orientation of the hyperplane.

SVM can handle both linearly separable and non-linearly separable data by using different kernel functions, such as linear, polynomial, and radial basis function (RBF) kernels. The default choice of kernels, RBF is used in this study. One of the strengths of SVM is its ability to handle high-dimensional data and small sample sizes, which makes it

suitable for a wide range of applications. It can also handle imbalanced datasets by adjusting the class weights or using cost-sensitive learning. Another advantage of SVM is its ability to provide a sparse solution, meaning that only a subset of the examples (i.e., the support vectors) are used to define the hyperplane, which can improve its scalability and interpretability.

5) *AdaBoost Classifier (ADB)*: AdaBoost Classifier (ADB) is a machine learning algorithm that belongs to the family of ensemble methods. It is a meta-algorithm that combines multiple weak classifiers to form a strong classifier. The basic idea behind AdaBoost is to iteratively train a series of weak classifiers on weighted versions of the training data, and then combine their predictions to obtain the final classification [46,47]. At each iteration, AdaBoost adjusts the weights of the training examples to give more importance to the misclassified examples. The weak classifier is then trained on the weighted data and added to the ensemble with a weight that depends on its classification accuracy. The weights of the examples are then updated again based on the classification errors of the current ensemble, and the process is repeated for a predefined number of iterations or until a desired accuracy is achieved. The base estimator parameter was set to Decision Tree Classifier with max depth = 1 which is the default. The number of estimators parameter was set to 50 by default.

One of the strengths of AdaBoost is its ability to improve the performance of weak classifiers and handle complex decision boundaries. It can also handle imbalanced datasets and noisy data, by adjusting the weights of the examples and focusing on the difficult examples. Another advantage of AdaBoost is its simplicity and interpretability, as it combines simple and easily understandable models to form a more complex and accurate model.

6) *Random Forest Classifier (RFC)*: Random Forest Classifier (RFC) is a machine learning algorithm that belongs to the family of ensemble methods. It is a meta-algorithm that combines multiple decision trees to form a powerful and robust classifier. The basic idea behind Random Forest is to build a collection of decision trees on randomly selected subsets of the training data, and then combine their predictions to obtain the final classification [48]. At each iteration, Random Forest randomly selects a subset of the features and a subset of the training features, and, trains a decision tree on the selected data. The trees are grown to maximum depth and may be pruned later to improve their generalization performance. The predictions of the individual trees are then combined using a majority vote or weighted vote, depending on the specific implementation. In this study, the number of estimators in RFC which denotes the number of trees in the forest was set to 100. The random state parameter was set to 0 which controls both the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking

for the best split at each node.

One of the strengths of Random Forest is its ability to handle complex decision boundaries and nonlinear relationships between the features and the target variable. It can also handle missing data, outliers, and irrelevant features, by randomly selecting subsets of the features. Another advantage of Random Forest is its ability to provide estimates of the importance of the features, which can help in feature selection and understanding the underlying data. It can also be easily parallelized and scaled to large datasets and distributed computing environments.

IV. EVALUATION

The evaluation of the machine learning models plays a crucial role in assessing the performance of the classification models. Employing appropriate evaluation settings and metrics, analyzing performance scores, and, identifying key findings can draw meaningful conclusions about the model's effectiveness and its relevance to the research objectives. The evaluation section forms an essential component of the study, contributing to the overall understanding and interpretation of the classification model's performance. This section provides insights into the model's performance by employing evaluation settings and metrics. It also discusses the obtained performance scores, key findings, and, the answer to the research question posed in this study.

A. Experiment settings

Machine learning classification models are evaluated using various settings and metrics to assess their performance and effectiveness. In this study, 6 machine learning models were evaluated in a 10-fold cross-validation setting. The total number of records present in the dataset is 5686. During the process, 5118 records were used in the training process and 586 records were used for the evaluation at each of the 10-folds. At every iteration, the data were randomly shuffled.

B. Evaluation metrics

The models were evaluated with the most common classification evaluation metrics used by scholars [49]. The evaluation metrics used in this study are accuracy (A), precision (P), recall (R), and, f1-score (F). Classification accuracy is a performance metric used to measure the accuracy of a machine learning model in predicting categorical outcomes [50]. It is the proportion of correct predictions made by the model out of all the predictions made. Classification accuracy of the developed models on unseen data is one of the principal metrics for evaluating classification models. The following formula has been used to determine classification accuracy:

$$\text{Accuracy (A)} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

In machine learning classification tasks, precision is another important performance metric [50] that measures the ability of

a classification model to identify only the relevant instances of a particular class in a particular dataset. More specifically, precision is the proportion of true positives (i.e., instances that are correctly classified as belonging to a certain class) out of all the instances that the model classified as belonging to that class (i.e., true positives plus false positives). In mathematical terms, it can be expressed as:

$$\text{Precision (P)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

Precision is an important metric in many real-world applications, especially to avoid falsely labeling instances as belonging to a certain class. It is generally used in conjunction with recall to get a better understanding of the model's performance. Recall is another important performance metric in machine learning classification tasks [50]. It measures the ability of a classification model to identify all relevant instances of a particular class in the dataset. More specifically, recall is the proportion of true positives (i.e., instances that belong to a certain class and are correctly classified as such) out of all the instances that actually belong to that class (i.e., true positives plus false negatives). In mathematical terms, it can be expressed as:

$$\text{Recall (R)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

F1-score is another performance metric that combines both precision and recall into a single metric. It provides a way to balance the two metrics and evaluate the overall performance of a classification model. The F1-score is calculated as the harmonic mean of precision and recall, giving equal weight to both measures. The harmonic mean balances the impact of precision and recall, ensuring that the F1-score remains high only if both precision and recall are high. The F1-score is calculated as:

$$\text{F1-score (F)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

C. Model Performances

Table II shows the combined minimum and maximum Accuracy (A), F1-score (F), Precision (P), and, Recall (R) obtained by the machine learning models over 10-fold cross-validation process for the 16 target variables. The target features are placed in the left-most column followed by the minimum and maximum results obtained in all 4 evaluation metrics by the classification models. These results span over all 10-folds and 6 machine learning models, thus a larger gap between the minimum and maximum values can be observed. The weighted average has been used in the calculation of precision, recall, and, F1-score as this study deals with multi-class classification tasks. Precision measures the accuracy of positive predictions, while recall measures the completeness of positive predictions. High precision and high recall are desirable, but there may be a trade-off between the two metrics in some cases. Precision

TABLE II
ACCURACY (A), F1-SCORE (F), PRECISION (P), AND, RECALL (R)
OBTAINED ACROSS THE TARGET VARIABLES

Metrics Target	A		F		P		R	
	Min	Max	Min	Max	Min	Max	Min	Max
ATHOME	0.41	0.60	0.38	0.51	0.31	0.57	0.40	0.60
EDUCATION	0.25	0.45	0.17	0.41	0.11	0.48	0.25	0.45
EMPLOYHH	0.46	0.75	0.31	0.72	0.23	0.76	0.46	0.75
EQUIPMUSE	0.33	0.69	0.21	0.68	0.14	0.69	0.33	0.69
HHAGE	0.38	0.65	0.21	0.64	0.15	0.68	0.38	0.65
MONEYPY	0.19	0.37	0.08	0.35	0.05	0.35	0.19	0.37
NHSLDMEM	0.37	1.00	0.20	1.00	0.13	1.00	0.37	1.00
NUMADULT	0.54	1.00	0.38	1.00	0.30	1.00	0.54	1.00
NUMCHILD	0.65	1.00	0.54	1.00	0.46	1.00	0.65	1.00
TEMPGONE	0.30	0.61	0.20	0.59	0.13	0.59	0.30	0.61
TEMPGONEAC	0.29	0.67	0.15	0.68	0.10	0.69	0.29	0.68
TEMPHOME	0.33	0.65	0.24	0.63	0.17	0.65	0.33	0.65
TEMPHOMEAC	0.29	0.69	0.13	0.68	0.09	0.70	0.29	0.69
TEMPNITE	0.31	0.62	0.18	0.61	0.12	0.64	0.31	0.62
TEMPNITEAC	0.28	0.66	0.15	0.65	0.10	0.67	0.28	0.66
USEWWAC	0.71	0.89	0.64	0.87	0.57	0.87	0.71	0.89

TABLE III
AVERAGE ACCURACY OF THE MACHINE LEARNING MODELS OVER
10-FOLD VALIDATION PROCESS

Models	ADB	CART	KNN	LDA	RFC	SVM	Mean
ATHOME	0.58	0.43	0.43	0.54	0.57	0.55	0.52
EDUCATION	0.38	0.3	0.28	0.40	0.40	0.33	0.35
EMPLOYHH	0.71	0.61	0.48	0.70	0.71	0.48	0.62
EQUIPMUSE	0.55	0.61	0.37	0.60	0.66	0.38	0.53
HHAGE	0.59	0.52	0.40	0.59	0.61	0.38	0.52
MONEYPY	0.34	0.26	0.21	0.31	0.35	0.22	0.28
NHSLDMEM	0.37	0.99	0.53	0.96	0.76	0.37	0.66
NUMADULT	0.77	0.99	0.64	0.96	0.77	0.54	0.78
NUMCHILD	0.69	0.99	0.67	0.97	0.75	0.68	0.79
TEMPGONE	0.56	0.51	0.32	0.51	0.55	0.37	0.47
TEMPGONEAC	0.47	0.63	0.31	0.58	0.64	0.31	0.49
TEMPHOME	0.55	0.60	0.37	0.55	0.63	0.41	0.52
TEMPHOMEAC	0.56	0.66	0.32	0.62	0.67	0.30	0.52
TEMPNITE	0.58	0.54	0.33	0.50	0.58	0.35	0.48
TEMPNITEAC	0.57	0.61	0.30	0.53	0.63	0.31	0.49
USEWWAC	0.86	0.85	0.74	0.83	0.88	0.76	0.82

evaluates the correctness of positive predictions, whereas recall assesses the inclusiveness of positive predictions. Ideally, both high precision and high recall are preferred; however, there can be instances where there is a trade-off between these two metrics. The F1-score is a measure of a model's accuracy that considers both precision and recall. The maximum accuracy, precision, recall, and, F1-score of 100% was observed in the case of the 3 target variables indicating the number of household members (NHSLDMEM), number of adults (NUMADULT), and, number of children (NUMCHILD) in the house indicating the machine learning models performed best in classifying these characteristics. The lowest accuracy, precision, recall, and, F1-score were obtained for the gross household income (MONEYPY) which yields the machine learning models didn't perform well in classifying this particular feature. It also indicates that the gross income of a household doesn't rely on the selected input features of this study. Low precision of 10% was observed in classifying average room temperature during summer nights (TEMPNITEAC) and summer days when no one is at home (TEMPGONEAC). The highest precision for these features are 67% and 69% respectively. This indicates that

some models performed poorly in classifying these two target features however some models performed above average in classifying these characteristics as there are 7 classes of data in each of these features.

Table III reveals the average accuracy obtained by each classification model during the 10-fold cross-validation process over the 16 target variables. The targets are positioned in the left-most column, followed by the average accuracy obtained by the six classification models. The right-most column shows the mean accuracy obtained for the respective target feature. The best average score for each target is highlighted in bold numbers. Across all 16 target variables, CART, LDA, and, RFC classifiers achieved the highest average accuracy (63%). KNN and SVM achieved the lowest average accuracy (42%). CART and LDA achieved over 95% average accuracy for 3 target variables which are the number of household members (NHSLDMEM), number of adults (NUMADULT), and, number of children (NUMCHILD). Most of the models also obtained above 75% accuracy for classifying main cooling equipment control behavior (USEWWAC). The average accuracy among all the models for USEWWAC is 82% meaning the models performed moderately well in classifying this feature. The models achieved more than 50% accuracy for all target variables except 2 which are the level of education (EDUCATION) and the total household income (MONEYPY) where the average accuracy obtained by the models are 35% and 28% respectively. The data in MONEYPY is distributed in 16 classes which is almost double compared to the other targets. This might answer the lower accuracy, however, the data in EDUCATION is distributed in only 5 classes and this does not support the previous claim. A more concrete reason can be the level of education and gross household income may not depend on the energy usage and housing detail input variables. Here, education refers to the education of individual occupants (RECS survey respondents) and MONEYPY is the gross household income. It is plausible that these two characteristics are independent of the chosen input variables in this study.

D. Answer to the research questions

Analysis of the results obtained by the machine learning models provides multiple insights that can help answer the research questions poised in section I. These findings can elaborate on the effectiveness of using ML models in the task of predicting building occupant characteristics.

RQ1: How effectively can machine learning tools predict individual building occupant characteristics?

Understanding the effectiveness of using machine learning tools in predicting individual building occupant characteristics is crucial as it explains the usability and feasibility of the task. The evaluation results can benefit in understanding

TABLE IV
AVERAGE PRECISION (P), RECALL (R), AND, F1-SCORES (F) OF CLASSIFICATION MODELS

Models	ADB			CART			KNN			LDA			RFC			SVM		
Target	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ATHOME	0.45	0.58	0.49	0.45	0.44	0.44	0.36	0.43	0.39	0.48	0.54	0.49	0.42	0.57	0.45	0.31	0.55	0.39
EDUCATION	0.39	0.38	0.38	0.31	0.31	0.31	0.27	0.28	0.27	0.40	0.40	0.35	0.41	0.40	0.37	0.11	0.33	0.17
EMPLOYHH	0.68	0.71	0.68	0.62	0.62	0.61	0.44	0.48	0.46	0.63	0.70	0.66	0.65	0.71	0.67	0.23	0.48	0.31
EQUIPMUSE	0.47	0.55	0.48	0.62	0.61	0.61	0.34	0.37	0.34	0.60	0.60	0.58	0.66	0.66	0.65	0.14	0.38	0.21
HHAGE	0.60	0.59	0.59	0.52	0.52	0.52	0.39	0.40	0.39	0.64	0.59	0.55	0.62	0.61	0.59	0.15	0.38	0.21
MONEYPY	0.31	0.34	0.32	0.26	0.26	0.26	0.19	0.21	0.20	0.32	0.31	0.29	0.31	0.35	0.31	0.05	0.22	0.08
NHSLDMEM	0.14	0.37	0.20	0.99	0.99	0.99	0.50	0.53	0.51	0.96	0.96	0.95	0.71	0.76	0.72	0.14	0.37	0.20
NUMADULT	0.69	0.77	0.71	0.99	0.99	0.99	0.59	0.64	0.60	0.95	0.96	0.96	0.72	0.77	0.71	0.30	0.54	0.38
NUMCHILD	0.68	0.69	0.66	0.99	0.99	0.99	0.59	0.67	0.62	0.96	0.97	0.96	0.68	0.75	0.69	0.46	0.68	0.55
TEMPGONE	0.55	0.56	0.55	0.52	0.52	0.52	0.29	0.32	0.30	0.51	0.51	0.49	0.54	0.55	0.50	0.13	0.37	0.20
TEMPGONEAC	0.40	0.47	0.38	0.63	0.63	0.63	0.31	0.31	0.31	0.56	0.58	0.53	0.62	0.64	0.61	0.10	0.31	0.15
TEPHOME	0.53	0.55	0.52	0.60	0.59	0.59	0.33	0.37	0.34	0.56	0.55	0.52	0.62	0.63	0.60	0.17	0.41	0.24
TEPHOMEAC	0.54	0.56	0.54	0.66	0.66	0.67	0.30	0.32	0.30	0.65	0.62	0.59	0.66	0.67	0.64	0.09	0.30	0.14
TEMPNITE	0.57	0.58	0.57	0.55	0.54	0.55	0.30	0.33	0.31	0.50	0.50	0.45	0.58	0.58	0.52	0.12	0.35	0.18
TEMPNITEAC	0.59	0.57	0.57	0.61	0.61	0.61	0.28	0.30	0.28	0.53	0.53	0.47	0.63	0.63	0.61	0.10	0.31	0.15
USEWWAC	0.80	0.86	0.82	0.85	0.85	0.85	0.62	0.74	0.66	0.86	0.83	0.83	0.85	0.88	0.86	0.57	0.76	0.65

the effectiveness of using the machine learning tools of this study. This work incorporates 6 state-of-the-art machine learning models for 16 selected occupant characteristics classification and prediction tasks. From the performance measures presented in table II, it can be deduced that the machine learning models performed efficiently in the task. The models performed significantly well for 4 occupant characteristics and moderately well for 10 characteristics. The models obtained below-average accuracy in classifying 2 characteristics (education and income) which in itself is a finding denoting these characteristics may not depend on the energy usage and selected housing features.

From table II it is clear that the performance of the machine learning models varied over the 16 target variables. The ML models obtained different accuracy, precision, recall, and, f1-scores for the different occupant characteristics. The 16 target attributes of this work can be separated into 3 categories, (i) Temperature preference and equipment control, (ii) Household characteristics, and, (iii) Individual characteristics. The ML models achieved different effectiveness in these 3 categories. The following is a description of the machine learning model performances over these categories:

(i) Temperature preference and equipment control: Targets 1 through 8 of table I fall in this category. It includes temperature preferences during winter and summer along with the usage of heating and cooling equipments. These attributes represent the entire household but are specific to temperature control. The temperature preference attributes achieved an average accuracy of 50%. The data in these variables are distributed in 7 categories which makes 50% a moderately good number for accuracy. The accuracy scores among these variables also do not vary more than 3 which makes sense with the same input features. Surprisingly, the classification of the usage of the main cooling equipment during the summer (USEWWAC) achieved significantly better average accuracy (82%) than the classification of the usage of

the main heating equipment (EQUIPMUSE) during the winter (52%). The random forest classifier (RFC) achieved highest accuracy in both of these target labels (88% for USEWWAC and 66% for EQUIPMUSE). This result indicates that the input features may provide more correlating information towards cooling equipment usage than heating equipment usage.

(ii) Household characteristics: Targets 12 to 16 of table I fall in this category. It includes the number of household members, number of adults, number of children, number of days, someone present at home, and, the combined household income. These attributes represent the entire household. The target attributes related to the number of people performed surprisingly well. The decision tree classifier achieved 99% accuracy in all three. It can be deduced that these variables have a high correlation with the input features present in the dataset. The input features like energy consumption, and, housing structure like number of bedrooms, total area, etc. can tell a lot about the number of habitats present. However, the machine learning models did not perform well in classifying the number of weekdays some is present at home (ATHOME) and the combined household income (MONEYPY) achieving an average accuracy of 51% and 28% respectively. The random forest classifier achieved the highest accuracy in these two variables (57% and 35% respectively). It can be the case that these variables are independent of the input features present in the dataset.

(iii) Individual characteristics: Targets 9 to 11 of table I fall in this category. It includes the age, employment status, and, level of education of the respondent. These attributes represent the individual who responded to the survey questionnaire. The machine learning models achieved an average accuracy of 51% for the respondent's age, 61% for the employment status of the respondent, and, 35% for the level of education of the respondent. These results indicate that the correlating input features were not enough for the machine learning model

to classify these target attributes. These variables are solely about independent respondents and not the entire household however most of the input variables in this work represent the entire housing unit. This makes it difficult for machine learning models to classify individual occupant characteristics from household data.

RQ2: How do machine learning algorithms compare with each other when predicting building occupant characteristics?

The comparison of the machine learning algorithms for any classification or prediction task is very important as it facilitates performance evaluation, model selection, optimization, interpretability, and, understanding of model behavior. It helps to make informed decisions, improve performance, and gain insights into the classification task and the models' characteristics. For comparing machine learning classification models, factors like accuracy, precision, recall, and, F1-score are typically considered to assess the performance and determine which model is better suited for the task. The models used in this study are the most commonly used machine learning classification models used by scholars [40, 50]. Even though the same input and output features were provided to each model, the evaluation results significantly varied. Analyzing table III and table IV can provide insights on the performance comparison among the machine learning tools used in this study.

Table III shows the average Accuracy obtained by each model. The reason behind the success of RFC can be the resourcefulness for making accurate predictions needed in strategic decision-making. Overall, the random forest classifier (RFC), the decision tree classifier (CART), and, the linear discriminant analysis classifier (LDA) are tied in the first position with an average accuracy of 63%. However, RFC outperformed the rest by achieving the best accuracy for 9 out of 16 target variables. The Adaboost classifier (ADB) can be placed in the second position as it achieved the best accuracy for 4 out of 16 target variables. The decision tree classifier achieved the highest accuracy of 99% for 3 of the target variables. Both K nearest neighbor classifier (KNN) and support vector machine classifier (SVM) scored an average of 42% overall accuracy.

Table IV shows the average Precision (P), Recall (R), and, F1-score (F) obtained by the six machine learning models throughout the 10-fold validation process across the 16 target variables. The target variables are placed in the left-most column, followed by average P, R, and, F obtained by each individual classification model. The results of each model are separated by vertical lines. Precision measures the accuracy of the positive predictions. Higher precision values indicate a lower false positive rate. For most target variables, RFC and CART have relatively high precision values, indicating their ability to accurately classify positive instances. Recall measures the ability of the model to identify positive instances

correctly. Higher recall values indicate a lower false negative rate. The RFC and CART tend to have higher recall values, closely followed by LDA for most target variables, indicating their ability to capture a higher percentage of positive instances. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. For most target variables, the RFC algorithm achieved higher F1-scores, indicating a good balance between precision and recall. Different algorithms performed better or worse depending on the target variables. For example, the CART and LDA tend to perform well for the population attributes while ADB and RFC performed relatively better in classifying education and gross household income target variables. This suggests that the performance and predictability of the classification model vary depending on the specific target variable being predicted. In summary, the table highlights the importance of selecting the appropriate machine learning model for a given classification task, as well as the potential for variation in performance across different targets.

V. DISCUSSION

The classification and prediction of occupant characteristics using machine learning tools can be useful in many ways. It can provide a way of partially automating the building occupant persona generation procedure which can facilitate both the occupants and the building designers in many ways. The following is an example case study of occupant persona generated from the feature predictions from this work followed by a discussion on the shortcomings of this study along with the possible future research directions.

A. Sample Occupant Persona Prediction

Table V shows a sample of prediction data from the machine learning models involving all 16 occupant characteristics. A random row from the RECS-2015 was provided as input to the ML models which predicted the 16 features. Table V reveals predictions from models with the best average accuracy (shown in table III) for each target feature. The prediction column of table V shows the actual prediction value from the models (numerical). The description column provides the textual description of each prediction from the class details of table I. OpenAI's ChatGPT [51] was utilized with the information present in table V for generating a descriptive occupant persona that yielded the following output:

"The individual is a middle-aged adult, between the ages of 31 and 50. The residents are employed full-time and have attained some college education or an associate degree. Their household consists of two adults, with no children residing in the house. The gross household income falls within the range of approximately \$60,000 to \$74,999. This income level can help inform decisions related to energy consumption, budgeting, and overall financial considerations.

During the winter, the heating temperature is set and left at a consistent level most of the time. When people are at home

TABLE V
SAMPLE OCCUPANT CHARACTERISTICS PREDICTION

Feature	Prediction	Description
EQUIPMUSE	1	Set a heating temperature and leave it there most of the time during the winter
USEWWAC	1	Set a cooling temperature and leave it there most of the time during the summer
TEMPHOME	4	Typical temperature during winter days when people are at home is 65 - 68 °F
TEMPGONE	2	Typical temperature during winter days when no one at home is 51 - 60 °F
TEMPNITE	4	Typical temperature during winter nights when people are at home is 65 - 68 °F
TEMPHOMEAC	5	Typical temperature during summer days when people are at home is 69 - 72 °F
TEMPGONEAC	7	Typical temperature during summer days when no one at home is 77 - 96 °F
TEMPNITEAC	4	Typical temperature during summer nights when people are at home is 65 - 68 °F
HHAGE	3	Middle age adult (31 -50)
EMPLOYHH	1	Employed full time
EDUCATION	3	Some college or associate degree
NHSLDMEM	2	2 People live in the house
NUMADULT	2	2 Adults live in the house
NUMCHILD	0	There are no children living in the house
ATHOME	0	Noone is present at home during the weekdays
MONEYPY	13	Gross household income is about \$60,000 to \$74,999

during winter days, the typical temperature is maintained between 65 and 68 °F, providing a comfortable and cozy environment. However, when no one is at home during winter days, the temperature is adjusted to a lower range of 51 to 60 °F to conserve energy and reduce heating costs. During winter nights when people are at home, the temperature is again set between 65 and 68 °F to ensure a pleasant sleep environment.

In the summer, a cooling temperature is set and maintained throughout most of the time. When people are at home during summer days, the typical temperature is kept between 69 and 72 °F, providing a cool and comfortable atmosphere. However, when no one is present at home during summer days, the temperature is allowed to rise to a higher range of 77 to 96 °F, conserving energy and minimizing cooling expenses. During summer nights when people are at home, the temperature is adjusted back to the comfortable range of 65 to 68 °F for a restful sleep.

Considering their employment, the household members are likely absent from home during weekdays. This schedule suggests that the heating or cooling systems may be adjusted accordingly to conserve energy and optimize comfort during those periods.”

The above is the ChatGPT-generated building occupant persona that harnessed 16 occupant characteristics predicted by the best-performing machine learning models of this study. This is a representation of a composite person who is likely to occupy a particular building or space. It can help designers and developers to better understand and empathize with their target audience, which can lead to better design decisions and a more user-centric approach to building design. It can act as a reference point throughout the design process and building designers can ensure that the building is tailored to the needs and preferences of its intended occupants. This demonstrates the possible usage and feasibility of using machine learning tools in the development of building occupant persona. It also

indicates the direction of automating the occupant persona development process.

B. Limitations and future research

The purpose of this research article was to investigate the application of state-of-the-art machine learning models in classifying building occupant characteristics. The ultimate objective was to automate the process of building occupant persona development, aiming to reduce manual effort and time consumption. The study successfully demonstrated the possibility of automating several steps in the generation of smart housing personas, thereby potentially streamlining the overall process. However, it is important to note that complete automation was not achieved in this study, leaving room for potential future research in this area. Further investigation is suggested focusing on fully automating the task, which would be a promising direction for future studies. This work utilized six different machine learning models in their analysis. Despite the extensive experimentation, this study couldn't distinguish any individual model that could effectively classify all 16 occupant characteristics with better confidence than the rest. Consequently, the adoption of a model selection strategy, where the best-performing models are chosen for each specific characteristic is suggested. This approach allows for more accurate predictions overall.

Future research works may explore deep-learning models for the classification and prediction of building occupant characteristics. Deep-learning techniques have recently shown superior performance in dedicated tasks compared to traditional machine-learning models. Therefore, incorporating deep-learning models into future research endeavors may yield improved accuracy and results. Additionally, the study proposes investigating the ensembling technique of machine learning models for occupant characteristic prediction. Ensemble learning combines the predictions of multiple models to enhance overall accuracy and robustness. Applying this approach to building occupant persona development could

potentially yield more reliable results. The research conducted in this study has practical implications for automating the persona development process in the context of building performance simulation, design behavior interventions, and smart building management solutions. By leveraging the findings of this research, future studies can work towards fully automating the persona development process. This holistic automation can lead to enhanced functionality and accuracy in various applications related to building performance and management.

VI. CONCLUSION

This research investigates the feasibility of using machine learning for the classification of occupant characteristics depending on housing and energy consumption data with a view of automating some steps of building persona development. In this study, six machine learning models have been utilized using the 2015 Residential Energy Consumption Survey data from the U.S. Energy Information Administration (EIA). The results indicate that it is possible to use machine learning tools for the classification and prediction of occupant characteristics which yielded the possibility of partially automating the process of building occupant persona development. This research contributes to the body of knowledge by proposing a machine learning-based approach for facilitating the development of building occupant personas. It proves that given the necessary data, it is possible to automatically predict the occupant characteristics with moderate confidence, and thus, help to better understand occupants and improve living conditions to meet occupant demands.

REFERENCES

- [1] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, and W. H. Engelmann, "The national human activity pattern survey (nhaps): a resource for assessing exposure to environmental pollutants," *Journal of Exposure Science & Environmental Epidemiology*, vol. 11, no. 3, pp. 231–252, 2001.
- [2] J. Bell, J. Mabb, V. Garcia Hansen, R. Bergman, and L. Morawska, "Occupant health and productivity: an australian perspective," in *Summary Book of The CIB 2003 International Conference on Smart and Sustainable Built Environment (SASBE 2003)*. Queensland University of Technology, 2003, pp. 1–8.
- [3] A. H. Buckman, M. Mayfield, and S. B. Beck, "What is a smart building?" *Smart and Sustainable Built Environment*, 2014.
- [4] P. Agee, X. Gao, F. Paige, A. McCoy, and B. Kleiner, "A human-centred approach to smart housing," *Building Research & Information*, vol. 49, no. 1, pp. 84–99, 2021.
- [5] E. Brangier and C. Bornet, "Persona: a method to produce representations focused on consumers' needs," in *Human factors and ergonomics in consumer product design*. CRC Press, 2011, pp. 37–61.
- [6] Z. Mo, "Intelligent buildings and intelligent agents-a human-centered framework for building controls," in *Proceedings of the 41st SICE Annual Conference. SICE 2002.*, vol. 5. IEEE, 2002, pp. 3151–3156.
- [7] H. S. Alavi, E. F. Churchill, M. Wiberg, D. Lalanne, P. Dalsgaard, A. Fatah gen Schieck, and Y. Rogers, "Introduction to human-building interaction (hbi) interfacing hci with architecture and urban design," pp. 1–10, 2019.
- [8] S. Takai and K. Ishii, "A use of subjective clustering to support affinity diagram results in customer needs analysis," *Concurrent Engineering*, vol. 18, no. 2, pp. 101–109, 2010.
- [9] S. M. H. Anik, X. Gao, N. Meng, P. R. Agee, and A. P. McCoy, "A cost-effective, scalable, and portable iot data infrastructure for indoor environment sensing," *Journal of Building Engineering*, p. 104027, 2022.
- [10] Y. Song, F. Mao, and Q. Liu, "Human comfort in indoor environment: a review on assessment criteria, data collection and data analysis methods," *IEEE Access*, vol. 7, pp. 119 774–119 786, 2019.
- [11] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.
- [12] T. M. Mitchell, "Artificial neural networks," *Machine learning*, vol. 45, pp. 81–127, 1997.
- [13] M. Pantic, "Introduction to machine learning & case-based reasoning," *London: Imperial College*, 2005.
- [14] J. Malik, R. Bardhan, T. Hong, and M. A. Piette, "Developing occupant archetypes within urban low-income housing: A case study in mumbai, india," in *Building Simulation*, vol. 15, no. 9. Springer, 2022, pp. 1661–1683.
- [15] W. O'Brien, I. Gaetani, S. Carlucci, P.-J. Hoes, and J. L. Hensen, "On occupant-centric building performance metrics," *Building and Environment*, vol. 122, pp. 373–385, 2017.
- [16] S. D'Oca, T. Hong, and J. Langevin, "The human dimensions of energy use in buildings: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 731–742, 2018.
- [17] S. Pan, X. Wang, Y. Wei, X. Zhang, C. Gal, G. Ren, D. Yan, Y. Shi, J. Wu, L. Xia *et al.*, "Cluster analysis for occupant-behavior based electricity load patterns in buildings: A case study in shanghai residences," in *Building simulation*, vol. 10. Springer, 2017, pp. 889–898.
- [18] K. Sun and T. Hong, "A framework for quantifying the impact of occupant behavior on energy savings of energy conservation measures," *Energy and Buildings*, vol. 146, pp. 383–396, 2017.
- [19] —, "A simulation approach to estimate energy savings potential of occupant behavior measures," *Energy and Buildings*, vol. 136, pp. 43–62, 2017.
- [20] B. Dong and K. P. Lam, "A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting," in *Building Simulation*, vol. 7. Springer, 2014, pp. 89–106.
- [21] B. Dong, Y. Liu, H. Fontenot, M. Ouf, M. Osman, A. Chong, S. Qin, F. Salim, H. Xue, D. Yan *et al.*, "Occupant behavior modeling methods for resilient building design, operation and policy at urban scale: A review," *Applied Energy*, vol. 293, p. 116856, 2021.
- [22] S. Hu, D. Yan, E. Azar, and F. Guo, "A systematic review of occupant behavior in building energy policy," *Building and Environment*, vol. 175, p. 106807, 2020.
- [23] M. A. Ortiz and P. M. Bluysen, "Developing home occupant archetypes: First results of mixed-methods study to understand occupant comfort behaviours and energy use in homes," *Building and Environment*, vol. 163, p. 106331, 2019.
- [24] G. Buttitta, W. J. Turner, O. Neu, and D. P. Finn, "Development of occupancy-integrated archetypes: Use of data mining clustering techniques to embed occupant behaviour profiles in archetypes," *Energy and Buildings*, vol. 198, pp. 84–99, 2019.
- [25] K. Amasyali and N. M. El-Gohary, "Real data-driven occupant-behavior optimization for reduced energy consumption and improved comfort," *Applied Energy*, vol. 302, p. 117276, 2021.
- [26] S. Carlucci, M. De Simone, S. K. Firth, M. B. Kjærgaard, R. Markovic, M. S. Rahaman, M. K. Annaqeeb, S. Biandrate, A. Das, J. W. Dziedzic *et al.*, "Modeling occupant behavior in buildings," *Building and Environment*, vol. 174, p. 106768, 2020.
- [27] X. Li and R. Yao, "A machine-learning-based approach to predict residential annual space heating and cooling loads considering occupant behaviour," *Energy*, vol. 212, p. 118676, 2020.
- [28] Z. Deng and Q. Chen, "Artificial neural network models using thermal sensations and occupants' behavior for predicting thermal comfort," *Energy and Buildings*, vol. 174, pp. 587–602, 2018.
- [29] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, "Using machine learning techniques for occupancy-prediction-based cooling control in office buildings," *Applied energy*, vol. 211, pp. 1343–1358, 2018.
- [30] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, and G. Brager, "Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning," *Building and Environment*, vol. 129, pp. 96–106, 2018.
- [31] Y. Sun, W. Hao, Y. Chen, and B. Liu, "Data-driven occupant-behavior analytics for residential buildings," *Energy*, vol. 206, p. 118100, 2020.
- [32] Z. J. Yu, F. Haghighat, B. C. Fung, E. Morofsky, and H. Yoshino, "A methodology for identifying and improving occupant behavior in residential buildings," *Energy*, vol. 36, no. 11, pp. 6596–6608, 2011.

- [33] M. Jia, R. S. Srinivasan, R. Ries, and G. Bharathy, "A framework of occupant behavior modeling and data sensing for improving building energy simulation," in *Proceedings of the Symposium on Simulation for Architecture and Urban Design*, 2018, pp. 1–8.
- [34] Y. Zhongguo, L. Hongqi, S. Ali, and A. Yile, "Choosing classification algorithms and its optimum parameters based on data set characteristics," *Journal of Computers*, vol. 28, no. 5, pp. 26–38, 2017.
- [35] U. E. I. Administration, *Residential Energy Consumption Survey Data 2015*, accessed May 31, 2022. [Online]. Available: <https://www.eia.gov/consumption/residential/data/2015>
- [36] A. Yarlagadda, J. Murthy, and M. K. Prasad, "A novel method for human age group classification based on correlation fractal dimension of facial edges," *Journal of king saud university-computer and information sciences*, vol. 27, no. 4, pp. 468–476, 2015.
- [37] Z. Lin, R. Yang, K. Li, G. Yi, Z. Li, J. Guo, Z. Zhang, P. Junxiang, Y. Liu, S. Qi *et al.*, "Establishment of age group classification for risk stratification in glioma patients," *BMC neurology*, vol. 20, no. 1, pp. 1–11, 2020.
- [38] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.
- [39] S. M. Weiss and I. Kapouleas, "An empirical comparison of pattern recognition, neural nets, and machine learning classification methods," in *IJCAI*, vol. 89. Citeseer, 1989, pp. 781–787.
- [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [41] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," in *Robust data mining*. Springer, 2013, pp. 27–33.
- [42] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 3. IEEE, 1996, pp. 1480–1483.
- [43] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.
- [44] Y. Yuan, L. Wu, and X. Zhang, "Gini-impurity index analysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3154–3169, 2021.
- [45] Y. Zhang, "Support vector machine classification algorithm and its application," in *International conference on information computing and applications*. Springer, 2012, pp. 179–186.
- [46] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, 1996.
- [47] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52.
- [48] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *International Conference on Intelligent Data Communication Technologies and Internet of Things*. Springer, 2018, pp. 758–763.
- [49] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.
- [50] A. Tharwat, "Classification assessment methods," *Applied computing and informatics*, vol. 17, no. 1, pp. 168–192, 2021.
- [51] OpenAI, "Chatgpt (version 3.5)," <https://chat.openai.com>, 2023.