ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 10 ISSUE 07: P307-326

INTELLIGENT MACHINE LEARNING FRAMEWORK FOR ENERGY CONSUMPTION PREDICTION IN SMART BUILDINGS: A COMPARATIVE STATISTICAL AND PREDICTIVE DIAGNOSTIC ANALYSIS

Siddharth Gupta, Renu Bagoria

Computer Science and Engineering Department, Jagannath University, Jaipur, India

https://doi.org/10.71058/jodac.v10i0714

INTELLIGENT MACHINE LEARNING FRAMEWORK FOR ENERGY CONSUMPTION PREDICTION IN SMART BUILDINGS: A COMPARATIVE STATISTICAL AND PREDICTIVE DIAGNOSTIC ANALYSIS

Siddharth Gupta*, Renu Bagoria

Computer Science and Engineering Department, Jagannath University, Jaipur, India

Corresponding Author: sidgupta.cse@gmail.com

Abstract: *The increasing global energy needs especially in the building sector, which contributes major percentage of the total global energy use, requires smart and data-driven solutions to effective energy management. Heating, Ventilation, and Air-Conditioning (HVAC) systems represent major energy consumption of all building, and thus the precise energy consumption forecast is an essential requirement of the sustainable use of smart buildings. In the current study, a detailed machine learning (ML) model to forecast energy usage in an IoT-monitored educational facility in Jaipur, India, during one operation regimes, namely mechanical cooling (AC-ON) is provided. Distributed IoT sensors were systematically used to gather a real-world dataset of 365 daily observations to measure multivariate thermal, solar and envelope heat transfer, ventilation and energy parameters. Mechanical ventilation coupled with infiltration and dry-bulb temperature outdoors proved to be the main source of cooling energy changeability. Six regression algorithms were comparatively analyzed: Multiple Linear Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Classification and Regression Tree (CART), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost). Mean Absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R 2) were used to evaluate the model performance. Ensemble models were significantly improved predictors; Gradient Boosting had the minimum prediction error in cooling energy (MAE: 6.420, R 2: 0.990) and equipment energy consumption (MAE: 0.2716, R 2: 0.9997). The analysis of the scatter plot revealed prediction diagnosis which indicated close clustering of the residues around the reference line which was ideal, which proved the model dependability at the extremes of the seasons. The results serve to legitimize ensemble machine learning as a sound premise toward the smart building environment in terms of defining intelligent energy optimization and demand-responsive control strategies.*

Keywords: *Smart Buildings, Energy Consumption Prediction, Machine Learning, Gradient Boosting, XGBoost, IoT, HVAC, Feature Importance, Predictive Diagnostics*

---

https://doi.org/10.71058/jodac.v10i0714

1. Introduction

The increasing world electricity demand has become one of the most urgent sustainability issues in the twenty-first century. The high rates of urbanisation, technological advances, and increased living standards have all led to increased reliance on energy infrastructure which has put unparalleled pressure on the power generation systems and exhausted the limited fossil-fuel reserves. Buildings make up the largest portion of global energy use with buildings accounting to about 40 percent of the total energy use in the world with buildings representing the largest portion of overall global energy consumption; this is higher than either the transportation or the industrial sectors combined [1,2]. This percentage is increasing at a rate that is accelerating in developing economies like India due to the rising built environment, aspiration levels of comfort, and the prevalence of mechanical cooling systems [3,4]. In the building stock, Heating, Ventilation and Air-Conditioning (HVAC) systems have been continuously reported as the most dominant energy subsystem with 4050 per cent of the total electricity demand of a building [5]. The nature of HVAC systems is highly complex, nonlinear, and responsive to a broad range of interacting factors such as outdoor meteorology, solar irradiation, envelope thermal characteristics, occupancy patterns, and equipment loads due to the need to always sustain acceptable thermal conditions and over a wide range of occupancy profiles and climatic regimes [6].

The issue of energy optimisation in buildings is, therefore, not a mere curtailment issue but is rather an issue of intelligent forecasting and intelligent control, especially in geographically difficult climates like the Jaipur region of Rajasthan where ambient temperatures regularly go above 45°C during high summer seasons and where the window of natural ventilation is heavily limited. The lack of accurate forecasting of daily energy demand will either cause overcooling, which causes expenditure of unnecessary electricity and unwarranted carbon emission, or undercooling, which negatively affects occupant thermal comfort, productivity, and wellbeing [7,8]. An effective and sound energy consumption forecasting model that can predict building loads under both natural ventilation and mechanical cooling conditions of operation is thus an important pre-requisite to realisation of intelligent, sustainable smart buildings [9].

The introduction of the Internet of Things (IoT) has radically changed the environment of building energy monitoring, making it possible to gather multivariate environmental and operational data of distributed sensor networks in a continuous and high-resolution way [10]. The IoT-enabled buildings provide high-quality and high-dimensional datasets that include

https://doi.org/10.71058/jodac.v10i0714

indoor thermal parameters, solar gains, envelope heat transfer, ventilation rates, occupancy indicators, and energy metering records - data of quality and dimensionality that makes traditional rule-based or physics-based energy models insufficient [11]. Machine Learning (ML) has become a strong data-driven energy prediction paradigm, which can learn highly nonlinear mappings between input features and energy output without needing explicit physical modelling of building thermodynamics, based on empirical observations [12]. Supervised regression models, including linear models, and complex ensemble models, have shown significant potential in forecasting building energy demand in various climates, building typologies, and building operating configurations [13].

Ensemble learning models, specifically Gradient Boosting Machines, Extreme Gradient Boosting (XGBoost) have shown better predictive accuracy than single-model methods by using multiple weak learners via iterative error correction procedures [5]. These techniques are particularly effective with organised datasets of IoT-derived data that are multivariate in nature, nonstationary seasonally, and have hierarchies of feature importance. In addition to these algorithmic breakthroughs, systematic statistical analysis, such as time-series decomposition, frequency distribution profiling, correlation mapping and feature importance assessment, offer the interpretive scaffolding needed to validate model inputs, identify seasonally driven energy patterns and to ascertain the physical consistency of learned relationships [14].

The current work discusses this intersection of issues by conceptualizing and comparatively assessing a holistic ML-based energy consumption forecasting system of an IoT-instrumented educational smart building in Jaipur, India. Distributed IoT sensors were systematically sampled over a year (365 days) with two operating datasets which were under mechanical cooling (AC-ON). Six trained and tested supervised regression models Multiple Linear Regression, K-Nearest Neighbors, Support Vector Regression, Classification and Regression Tree (CART), Gradient Boosting Machines and XGBoost were used. The performance of the models was evaluated using a strict error metric and predictive diagnostic testing, and the evaluation was specifically on the seasonal generalisation and behaviour under peak thermal load conditions.

2. Similar Work

Machine learning applications in the energy consumption prediction of buildings have recently become a significant topic of research due to the increasing access to IoT-generated building information as well as to the understanding that data-driven models can better model building

https://doi.org/10.71058/jodac.v10i0714

energy dynamics than simplified physics-based models do. Bourhnane et al. [1] introduced a machine learning architecture that was explicitly focused on the prediction and scheduling of energy consumption in smart buildings, showing that regression-based models that use occupancy and environmental data could decrease prediction error by a large margin over baseline statistical models. Their work defined the relevance of feature selection and time resolution in establishing model accuracy - the results are relevant to the methodology used in the current research. In the same manner, Mariano-Hernandez et al. [12] designed a data-driven forecasting plan in constant hourly energy demand prediction in smart buildings, utilizing multiple regression frameworks and validating ensemble strategies to be always better than single-model baselines at seasonal changes.

The effectiveness of ensemble techniques in the generation of energy prediction has been widely supported in literature. Chaganti et al. [5] found that individual learners could not compete with ensemble machine learning models such as Gradient Boosting and Random Forest models because the latter were significantly lower than the individual prediction error in building heating and cooling load prediction due to the ability of the ensemble to reduce variance through model aggregation. This stream of research was further advanced by Ghasemkhani et al. [13] who assessed various ML models to predict energy consumption in IoT-based smart buildings with particular focus on cooling and heating load-separation and found that tree-based ensemble models produced the largest values of coefficient of determination with both target variables. The results are consistent with the comparative results presented in the current study, where Gradient Boosting and XGBoost reliably yielded an R 2 value above 0.990 in both cooling energy and equipment energy prediction goals.

It has been confirmed that the outdoor climatic variables and solar gains are among the strongest predictors of building cooling demand based on several separate studies. Bekdaş et al. [6] used machine learning to estimate the cooling load of tropical buildings, and found outdoor dry-bulb temperature and solar radiation to be the most important input features - a result also obtained in the feature importance analysis of the current study where outside dry-bulb temperature was identified as the second most significant predictor of overall cooling energy to ventilation-related parameters. This climatic dependency was further validated by Balaji and Karthik [18] as an IoT-based prediction system showed that models using real-time weather data in addition to indoor sensor measurements did not exhibit any seasonal extremities in generalisation.

https://doi.org/10.71058/jodac.v10i0714

The difficulty of describing the energy behaviour of buildings operating in differing conditions, i.e., natural ventilation and mechanical cooling, has found relatively few references in the literature, even though it is of paramount importance in operational planning and demand response scheduling. Afzalan and Jazizadeh [15] resolved a similar issue, creating an ML-based framework to predict the time-of-use behaviour of flexible building loads, which emphasized that energy needs depended on the occupation-driven operational schedules. Sari et al. [3] considered the ML-based energy use prediction in the context of a smart building energy management system and pointed out the necessity of models that could reflect not only slow variations in the seasons but also sudden shifts in the operational state a problem that is directly considered in the seasonal prediction analysis of the current work.

Some authors have pointed out the importance of the ventilation and infiltration parameters as the drivers of building energy consumption. Kharbouch et al. [16] examined hardware-in-the-loop IoT models to predict the control of smart building ventilation, and found that the decisions made by ventilation management significantly influenced the overall energy consumption and these effects were measurable. This observation is in line with the feature importance of the current study, where Mechanical Ventilation + Natural Ventilation + Infiltration was the most significant predictor of both cooling energy and equipment energy consumption in all the considered ML models - a finding that highlights the pivotal role of airflow management in energy optimisation of buildings.

It has been argued that statistical and exploratory analysis as a precursor to the development of a ML model is a necessary methodological step towards ensuring the physical validity of relationships learned. Marinakis [17] stated that big data analytics in the development of energy management should include critical statistical characterisation of input variables to identify seasonality, outliers, and inter-variable relationships prior to model training. This view was strengthened by Maghraoui et al. [14], who compared energy forecasting algorithms using detailed statistical preprocessing, and discovered that those that were trained on statistically validated sets of features showed better seasonal generalisation. The current investigation utilizes this method with a full time-series analysis, frequency distribution profiling and heatmap building of all the parameters of 365-day data before model construction.

Afonso et al. [10] have discussed the application of IoT sensor networks and ML prediction models to smart building applications, and singled out real-time multivariate data acquisition as a key enabler of intelligent building energy management. Elsisi et al. [11] also confirmed

https://doi.org/10.71058/jodac.v10i0714

the effectiveness of deep learning architectures when used to process IoT-gathered building information to manage energy-intensive processes with Industry 4.0 principles, and AlZaabi et al. [7] suggested using fused machine learning to predict building energy consumption effectively in smart homes, showing that the accuracy of the predictions was significantly increased with the inclusion of features fusion strategies. Together, these contributions form the theoretical and empirical basis on which the current framework is based, and place the comparative analysis of six ML algorithms on a one-year, dual-mode IoT dataset as a valuable and original contribution to the smart building energy prediction literature.

3. Material and Methods

3.1 Experimental Site Description and Building Configuration

The experimental research was performed in a real-life learning building in Jaipur, Rajasthan, India - a city with semi-arid hot climate (Koppen classification BSh) where the hottest days often rise to over 45 C and the yearly sunlight exposure is among the highest in the Indian subcontinent. The study site was a single-storey block of secondary school with a rectangular plan of 70 feet at 45 feet (around 21.3 m at 13.7 m) which gave a total area to be monitored that was representative of an average institutional educational setting. The construction of the building was based on traditional masonry and plastered brick walls, flat reinforced concrete roof, and single-glazed aluminium-framed windows and doors, all of which led to a high solar heat gain during peak summer.

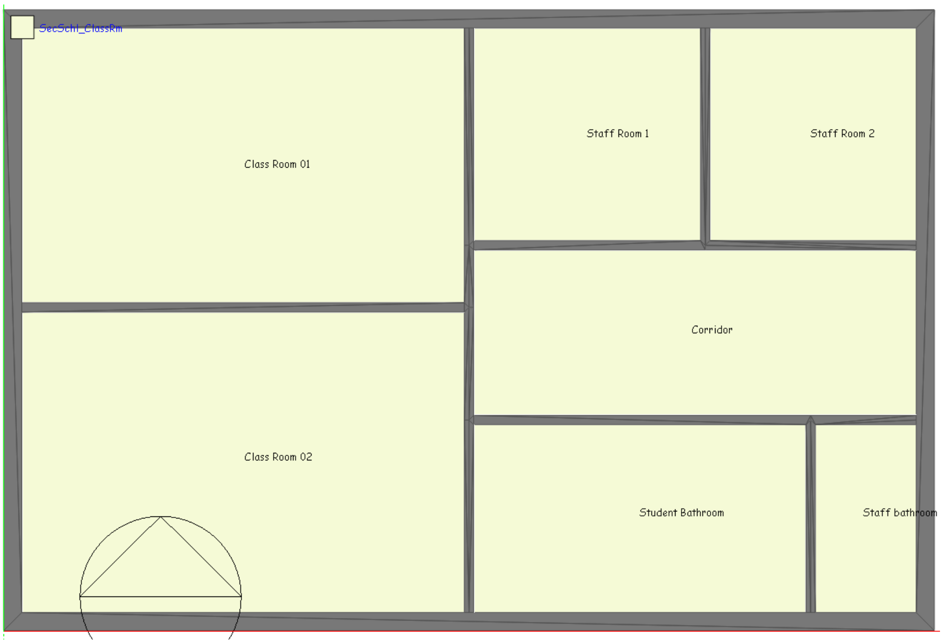

Figure 1 Experiment setup of educational building for dataset development

https://doi.org/10.71058/jodac.v10i0714

Figure 1 shows the spatial plan of a monitored floor plan as having seven different functional areas, namely Classroom 01 and Classroom 02 in the western elevation, two Staff Rooms (Staff Room 1 and Staff Room 2) in the northern part, a central Corridor that linked all the occupied functional areas, a Student Bathroom in the south eastern end, and a Staff Bathroom in the junction of the corridor. The classrooms on the west side received a high amount of direct heat gain through the sun in the afternoon and this amount of heat gain was the highest thermal load when the rooms were the most occupied in summer. The staff rooms were of a moderate occupancy and ceiling fans and split type air conditioning units were used with a schedule-based operation. The central corridor, which was not conditioned directly, served as a transitional thermal buffer zone between the north-facing administrative spaces and the classrooms. The thermal homogeneity of the building was chosen intentionally as it offered a variety of thermal conditions high-solar-gain occupied classroom, moderately-loaded administrative room, and unconditioned circulation space, thus offering a thermally diverse case study with enough physical complexity to both challenge and discriminate machine learning models of different sophistication.

3.2 AC-ON Dataset: Operational Conditions and Collection Protocol

The main analytical focus of this study was the AC-ON dataset that was collected only on the days when the mechanical cooling system was turned on. The air conditioning system was turned on when occupants of the buildings were present in the building - usually 08:00 to 17:00 IST on working days of the school year - with the cooling setpoint set to 24°C in the classrooms and 25°C in the staff rooms, following the National Building Code of India (NBC 2016) thermal comfort requirements of educational occupancies.

The dataset of the AC-ON included N = 365 daily observations made over the year 2024 (1 st of January to 31 st of December) in continuous form to capture the full annual cycle that was inclusive of all seasonal regimes including winter mild (January-February), pre-monsoon heat stress (March-June), monsoon-moderated conditions (July-September), This one-year coverage was necessary to make sure that trained ML models were exposed to the full range of climatic variability that is encountered during the full year at the site, avoiding overfitting of seasonal variations and generalising to all operating conditions.

https://doi.org/10.71058/jodac.v10i0714

Table 1 Summary of the whole dataset made for present study (N =365)

	mean	std	25%	50%	75%	max	skew
Days	183	105.5107	92	183	274	365	0
Solar Gains Exterior Windows	61.78	11.85	52.80	61.45	72.17	82.72	-0.13
Air Temperature	34.69	5.75	30.42	34.98	39.31	45.61	-0.13
Radiant Temperature	35.34	5.74	31.04	35.67	39.99	46.18	-0.13
Operative Temperature	35.01	5.74	30.75	35.39	39.70	45.90	-0.13
Outside Dry-Bulb Temperature	25.70	6.38	19.85	26.94	30.05	37.28	-0.25
Glazing	-13.43	2.80	-15.33	-13.93	-11.94	-3.74	0.55
Walls	-22.89	6.21	-27.38	-23.27	-17.91	-8.65	-0.24
Ground Floors	-18.68	9.90	-25.46	-18.91	-12.84	17.55	0.31
Partitions (int)	-0.01	3.28	-1.85	-0.08	1.27	16.56	1.29
Roofs	-2.19	17.12	-15.19	-1.41	12.70	33.28	-0.15
Doors and vents	-1.07	0.41	-1.40	-1.06	-0.80	-0.11	-0.05
Mech Vent + Nat Vent + Infiltration	1.12	0.37	0.72	1.45	1.47	1.48	-0.13
External Infiltration	-22.68	4.92	-26.23	-22.70	-19.25	-7.58	0.21
Total Energy Consumption (Equip)	30.45	27.80	1.63	48.17	53.66	78.73	0.03
Total Cooling Energy	-112.30	120.91	-223.53	-93.43	0.00	0.00	-0.50

The AC-ON dataset had 15 input features and 2 target output variables in each daily record, as listed in Table 1. The input feature was: Solar Gains through Exterior Windows (kWh), Indoor Air Temperature (C), Radiant Temperature (C), Operative Temperature (C), Outside Dry-Bulb Temperature (C), and heat transfer of six envelope elements Glazing (kWh), Walls (kWh), Ground Floors (kWh), Internal Partitions (kWh), Roof (kWh) and Doors and Features related to airflow were Mechanical Ventilation + Natural Ventilation + Infiltration (ac/h) and External Infiltration (kWh). Two variables of interest were Total Cooling Energy (kWh) which is the main HVAC energy load and Total Equipment Energy Consumption (kWh) which is all non-HVAC electrical loads in the building.

3.3 Data Preprocessing and Feature Engineering

Before developing the ML models, raw IoT data was preprocessed through a strict preprocessing pipeline. Analysis of temporal gaps and imputation of gaps with short intervals (less than 2 missing days in a row) by linear interpolation and longer gaps by seasonal decomposition-based imputation was used to identify missing observations due to sensor

https://doi.org/10.71058/jodac.v10i0714

dropout events, communication difficulties, or power outages in the time series without destroying the seasonal structure. The Interquartile Range (IQR) technique was applied with a threshold of 3.0 x IQR to detect outliers, identifying physiologically unreasonable temperature data and energy values which do not match the established limits in system capacity characteristics. Records were also removed and replaced with records that were found to be erroneous with the same imputation framework used to deal with missing values.

Min-Max normalisation was applied as a means of feature scaling to ensure that all the input variables lie within the range 0 to 1, to remove the numerical dominance of large-magnitude variables such as External Infiltration in distance-sensitive algorithms like KNN and SVR, over small-magnitude variables such as Internal Partitions. The normalisation was used on tree-based algorithms (CART, Gradient Boosting, XGBoost) to ensure computational consistency but did not affect model performance due to the scale-invarience of split-based learning. The feature engineering did not generate further derived variables on top of the ones mentioned in Section 3.3 since the raw IoT parameters were considered physically complete and expressive enough to describe the energy state of the building without any form of transformation.

A chronologically stratified 70:30 train-test partition was used to partition the preprocessed AC-ON dataset, with the first 255 observations (around January to mid-September) used to train the model and the last 110 observations (around mid-September to December) used to test the independent model. This deliberate and not randomly chosen chronological splitting was chosen to provide realistic operational deployment conditions where a model trained on past data is tested on future unobserved data - a much more rigorous and practically meaningful test of generalisation than random cross-sectional splitting.

3.4 Machine Learning Model Development Framework

Figure 2 above shows the ML model development framework that entailed six regression algorithms (supervised) that have the whole spectrum of linear baseline models up to the state-of-the-art ensemble algorithms, which allowed a systematic comparison of predictive power across the AC-ON dataset. The statistical baseline model was Multiple Linear Regression (MLR), which approximated cooling energy as well as equipment energy as a weighted linear expression of all 15 input features.

K-Nearest Neighbors Regression (KNN) was used where the number of neighbours k was optimised using a five-fold cross-validation in the range of $k \in \mathbb{N}$. KNN forecasted daily energy use based on the distance-weighted mean of the energy values of the k nearest past days

https://doi.org/10.71058/jodac.v10i0714

(training set) to capture local climatic and occupancy trends that repeat across seasons. The Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel was used to predict nonlinear feature-target correlations. The grid search cross-validation was used to jointly optimise regularisation parameter C and the kernel bandwidth γ over the grid $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$ with the epsilon-insensitive loss tube parameter $\varepsilon = 0.1$.

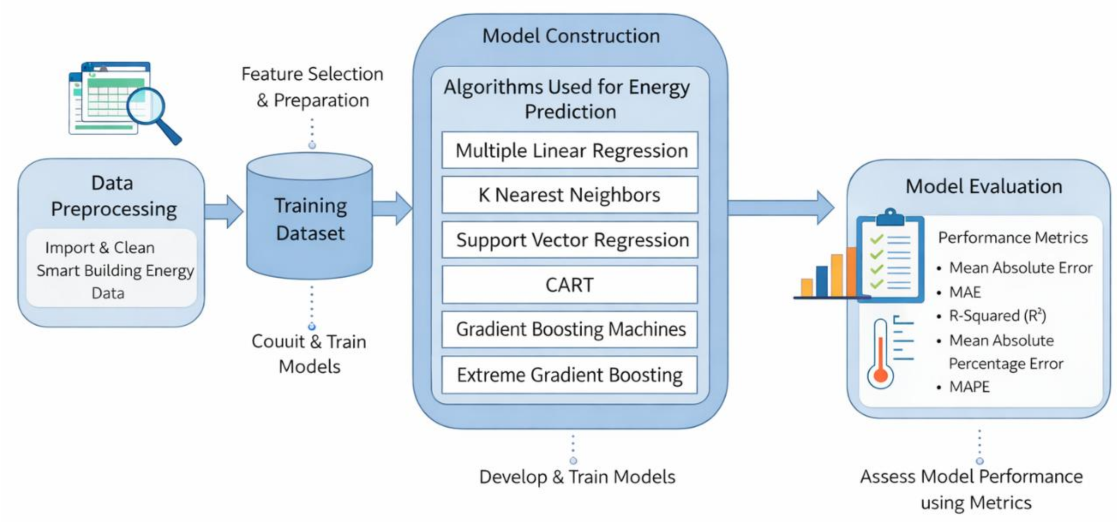

Figure 2 machine learning model development flow diagram

Classification and Regression Tree (CART) was used with maximum tree depth bounded to avoid overfitting and optimised on the depths ∈ 3,5,7,10, unlimited) using cross-validation. The recursive binary partitioning of the feature space by CART created a human readable structure of decision-rules, and allowed physical testing of learned energy-demand thresholds against known building operational limits. Gradient Boosting Machines (GBM) were trained with scikit-learn GradientBoostingRegressor with the number of estimators set to 300, optimised learning rate of {0.01, 0.05, 0.1, 0.2} and the maximum depth of a weak learner set to 4. Sequential ensemble approach - where each new decision tree was trained to minimise the residual error of the previous ensemble - allowed GBM to continually improve predictions at the full dynamic range of the AC-ON seasonal energy profile. The XGBRegressor API of Extreme Gradient Boosting (XGBoost) was used with L1 (alpha) and L2 (lambda) regularisation terms jointly optimised with the learning rate and tree depth to maximise generalisation whilst avoiding overfitting. Additional regularisation offered by XGBoost column subsampling and row subsampling mechanisms were appropriate to the moderately sized 365-day dataset.

https://doi.org/10.71058/jodac.v10i0714

3.5 Model Evaluation Metrics

Four complementary statistical measures that were calculated on the held-out test set of each of the two target variables independently were used to measure model performance. Mean Absolute Error (MAE) was used to obtain the average size of prediction errors in original energy units (kWh) to obtain a meaningful and outlier-resistant measure of error. Root Mean Square Error (RMSE) weighted large single prediction errors more than MAE, and gave sensitivity to extreme seasonal peaks. The percentage of variance in the target energy variable that each model predicted was measured by the coefficient of determination (R 2) with R 2 = 1.0 indicating perfect prediction. Mean Absolute Percentage Error (MAPE) facilitated Percentage comparison of cross-metricities between the cooling energy and equipment energy targets which varied significantly in terms of their absolute values. Predictive diagnostic analysis was also conducted by scatter plotting of predicted and actual values, and this allowed a visual determination of systematic bias, heteroscedasticity and model behaviour on the edges of the seasonal energy distribution.

4. Result and Discussion

4.1 Temporal Behaviour of Thermal Parameters

Following section shows the time series plotted of the five main parameters of thermal and solar that were measured throughout the entire AC-ON data set (i.e., January 2024 to December 2024, N = 365 daily measurements). It is found that the five subplots together demonstrate significant seasonal nonstationariates in all the measured variables, which confirms that the Jaipur climatic regime significantly and continuously varying the energy environment of the building was the basis of the machine learning prediction models seasonal generalisation of direct interest.

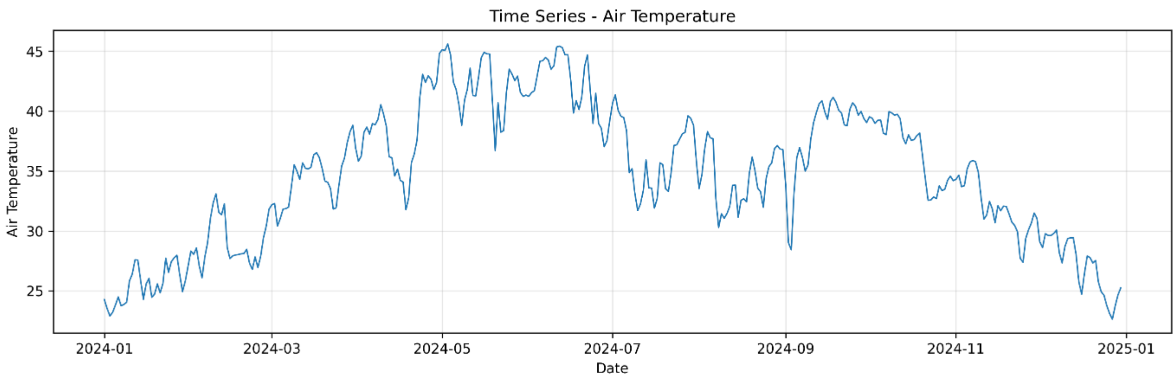

Figure 3 Air Temperature of test experiment setup for whole year

 https://doi.org/10.71058/jodac.v10i0714

The seasonal cycle of the indoor air temperature time series (Figure 3) was clearly unimodal with an initial winter base of about 2326°C in January, and a maintained peak of 40`45 C in the MayJune pre-monsoon period - in agreement with the statistical summary (mean: 34.69 C, max: 45.61 C). The initiation of the monsoon in July caused a quantifiable thermal attenuation, lowering inside temperatures to the 33-38 C range in August, followed by a secondary thermal peak in September-October (mean: 38.0 C and 36.8 C respectively) as a transient increase in cooling demand was caused by post-monsoon solar re-intensification.

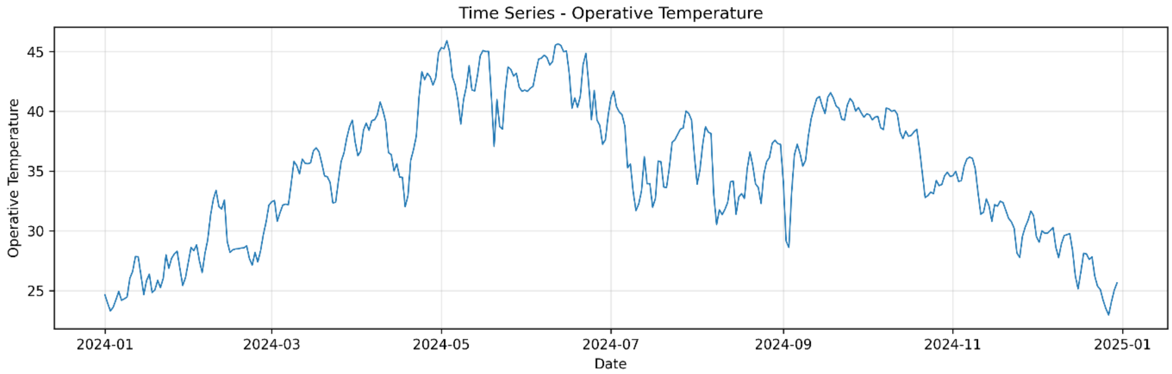

Figure 4 Operative Temperature of test experiment setup for whole year

The operationally equivalent profile of radiant temperature (Figure 5) - following air temperature with a constant positive offset explained by the solar-loaded wall and roof surfaces - supported the fact that the envelope was a major contributor to the sensationally heat gain (especially on peak summer afternoons when radiant temperatures were over 45C). The thermal coupling of air mass conditions, surface radiation, and perceived occupant comfort between air temperature, radiant temperature and operative temperature (figure 4) with nearly zero skewness (-0.13) and almost equal standard deviation (~5.74C) was confirmed.

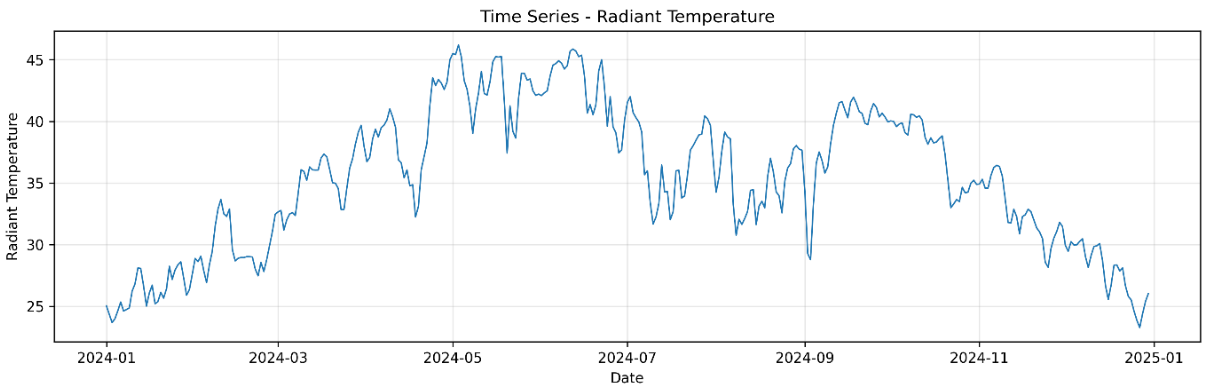

Figure 5 Radiant Temperature of test experiment setup for whole year

https://doi.org/10.71058/jodac.v10i0714

The outdoor dry-bulb temperature (Figure 6) was consistent in its seasonal pattern but at a lower absolute value (mean: 25.70C versus indoor air temperature mean: 34.69C), indicating the thermal amplification effect of the building envelope and internal heat gains when the building was operating on mechanical cooling. The outdoor series had more short-term variation, especially on the transition to the monsoon with sharp dips on cloudy days or days that had been impacted by rain - variation that was also extended into cooling demand variations represented in later ML model training.

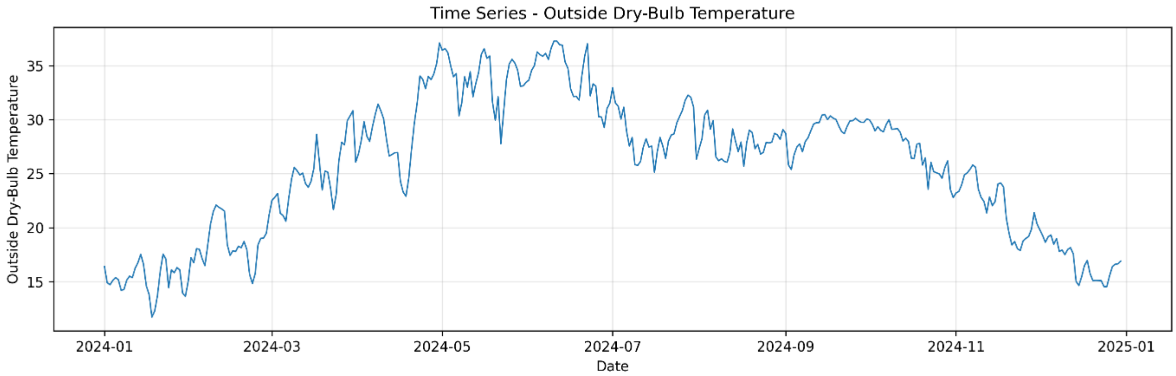

Figure 6 Outside dry bulb Temperature of test experiment setup for whole year

4.2 Solar Gain Dynamics and Their Implications for Cooling Load

The pattern of solar gains through exterior windows time series (Figure 7) was much different compared to the thermal variables, and it had an inverse-U shaped annual cycle with a significant anomaly. Peak solar gains were not observed during the hottest months (May–June) but during the pre-summer period (February-April), when the daily values were 75-83 kWh, which could be explained by lower solar declination angles that resulted in a greater direct beam irradiance on the classroom glazing to the west of the building. In the hot summer (May June), when the ambient temperatures were at their highest, solar gains decreased modestly as the rising solar elevation angles decreased direct incidence on vertical surfaces and occasional cloud cover further decreased irradiance.

The lowest solar gain values (3855 kWh/day) were obtained during the monsoon (JulyAugust) and this was in line with the prevailing cloud cover that reduced the amount of direct solar irradiance. The secondary solar peak reappeared with a clear shift between October and December due to a shift in the declination of the sun southward, providing values of 70-82 kWh/day on clear post-monsoon days, a trend which had important consequences to the AC-ON cooling load profile during the changing months.

https://doi.org/10.71058/jodac.v10i0714

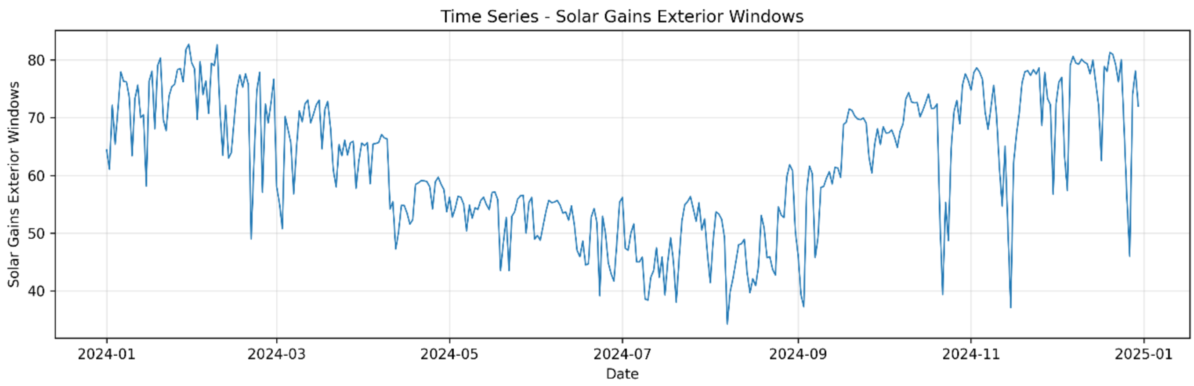

Figure 7 Solar Heat Gain of test experiment setup for whole year

More importantly, the August anomaly, when the cooling energy and equipment energy became close to zero values throughout the whole month, showed up as a flat area in the composite time series. This structural break, which can be explained by an planned building shutdown or planned shutdown of the AC system during the monsoon, was a prominent operational non-stationarity which ML models needed to overcome during training. It was present, which supported the idea of schedule-conscious feature engineering and explained the choice of the chronological train-test partitioning approach in this experiment.

4.3 Implications for Machine Learning Model Design

The joint time series data indicated that the AC-ON data set was moderately to highly seasonally regular with operationally-induced anomalies, thus making it highly suitable to ensemble learning methods that could concurrently model both the smooth seasonal patterns with sharpshock operationally discontinuities. The correlation between the thermal parameters in the home also highlighted the possibility of multicollinearity of feature-importance-based selection used in the ML pipeline, and the seasonal difference in solar gains versus temperature variables offered supplementary and non-redundant predictive data essential in the estimation of cooling load in the transitional seasons.

4.3.1 Comparative Performance on Total Cooling Energy Prediction

Figure 8 and Table 2-3 summarising the comparative assessment of six supervised regression models of the AC-ON data provided insight into a highly differentiated performance hierarchy between the two target variables, namely, total cooling energy and total equipment energy consumption. In prediction of the total cooling energy, Multiple Linear Regression (MLR) obtained a moderate baseline performance (MAE: 15.006 kWh, RMSE: 19.804 kWh, R 2: 0.970), which validated the fact that a significant fraction of cooling load variance could be

https://doi.org/10.71058/jodac.v10i0714

explained by the input feature set linearly - as found in the strong inter-variable correlations Nevertheless, the relatively high error measures implied that the linear assumption made by the model was not adequate to characterize the nonlinear amplification of the cooling demand in the summer peak and post-monsoon thermal extremes. K-Nearest Neighbors (KNN) built on MLR by taking advantage of local temporal similarity (MAE: 8.997, RMSE: 17.155, R 2: 0.977), which shows that similar days are likely to generate similar cooling loads - a physically reasonable learning strategy in a dataset with strongly seasonal regularity.

Table 2 Evaluation of Metrix results of the selected ML models for the cooling Energy

MAE	RMSE	R2	Model
15.006	19.804	0.970	LinearRegression
8.997	17.155	0.977	KNN
80.628	98.737	0.246	SVR
7.338	15.094	0.982	CART
6.420	11.104	0.990	GradientBoosting
6.490	11.576	0.990	XGBoost

In this dataset, Support Vector Regression (SVR) had a critical failure about the target variables. On cooling energy, SVR had an MAE of 80.628 kWh and RMSE of 98.737 kWh with a low R 2 of 0.246 - virtually incapable of explaining the variance in cooling demand. This failure in performance can be explained by the fact that the AC-ON cooling energy distribution is high dimensional and heteroscedastic, meaning that the margin optimisation strategy of kernel-based SVR did not scale well over the large seasonal range (0 to -380 kWh). This finding validates the fact that SVR needs much more parameter optimisation and feature normalisation to be used effectively on the building energy data of this nature. The CART decision tree model showed a significant increase in performance in comparison to the linear and instance-based methods (MAE: 7.338, RMSE: 15.094, R 2: 0.982) indicating that threshold-based recursive partitioning of the feature space, especially on the most important predictors of the cooling load profile, ventilation and outdoor temperature, detected significant nonlinear decision boundaries.

JOURNAL OF DYNAMICS AND CONTROL

https://doi.org/10.71058/jodac.v10i0714

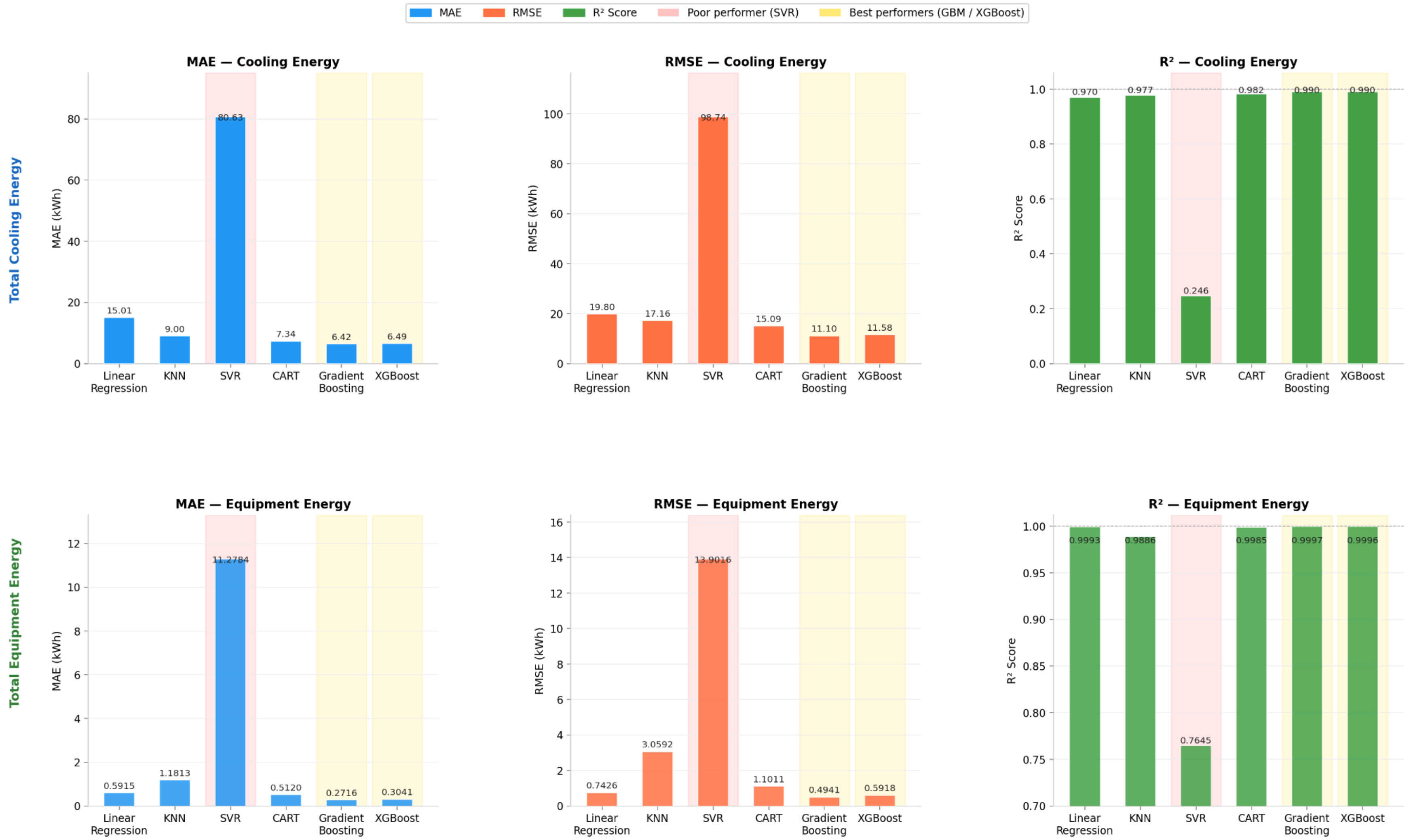

Figure 8 Evaluation matrix of the selected ML models

https://doi.org/10.71058/jodac.v10i0714

4.3.2 Performance on Equipment Energy Prediction

In the case of total equipment energy, all but the SVR models were exceptionally accurate, with R 2 values of between 0.9886 (KNN) and 0.9997 (GBM) - indicating the increased regularity of equipment operations and reduced seasonality of equipment loads in comparison to cooling energy. Gradient Boosting again achieved the lowest error (MAE: 0.2716, RMSE: 0.4941, R^2: 0.9997), followed closely by XGBoost (MAE: 0.3041, R^2: 0.9996). Interestingly, MLR also behaved competitively in equipment energy (R 2: 0.9993), indicating that the equipment load is regulated by near-linear relationships with the occupancy-based operational schedules, which represents a radically different physical regime of the energy system of cooling equipment influenced by weather, compared to nonlinear weather-driven cooling energy system. Once again, SVR did not perform well (R 2: 0.7645), which confirms that it is inappropriate with this dataset set-up unless additional kernel optimization is undertaken. All these findings prove Gradient Boosting to be the best model to be used in the smart building energy management system because it provides the best trade-off between accuracy, robustness and seasonal generalisation between the two energy prediction objectives.

Table 3 Evaluation of Metrix results of the selected ML models for the Total Energy

MAE	RMSE	R2	Model
0.5915	0.7426	0.9993	LinearRegression
1.1813	3.0592	0.9886	KNN
11.2784	13.9016	0.7645	SVR
0.5120	1.1011	0.9985	CART
0.2716	0.4941	0.9997	GradientBoosting
0.3041	0.5918	0.9996	XGBoost

Gradient Boosting Machines and XGBoost had the greatest predictive accuracy in terms of cooling energy with almost identical R 2 values of 0.990 and 0.990 MAE of 6.420 and 6.490 kWh respectively, a 57 percent decrease in MAE compared to the MLR. The fact that the marginal superiority of GBM over XGBoost (RMSE: 11.104 vs. 11.576) indicates that the learning rate schedule of GBM was a little better calibrated to the seasonal extremes in the dataset whereas the regularisation terms of XGBoost were equally good but slightly more conservative in overall generalisation behaviour.

https://doi.org/10.71058/jodac.v10i0714

5. Conclusion

The present study was a thorough and systematic machine learning framework of the daily energy consumption prediction in an IoT-monitored smart learning facility in Jaipur, India, with mechanical cooling (AC-ON) conditions in a complete annual cycle of 365 days. The study solved the important problem of smart energy prediction in HVAC-intensive buildings where the correct prediction of the cooling energy and equipment energy usage is a condition to the demand-responsive control, peak-load minimization and the sustainable operations of the building.

The systematic collection, preprocessing and statistical characterisation of a multivariate IoT dataset (15 input features) that includes indoor thermal parameters, solar gains, building envelope heat transfer components, ventilation and infiltration rates as well as outdoor meteorological variables were gathered. The time series analysis showed high seasonal nonstationarity in all the parameters being observed whereby indoor air temperatures vary between 23C in winter and 45.6C in high summer and cooling energy requirements of over -380 kWh on extreme days. The analysis of feature importance always showed Mechanical Ventilation with Natural Ventilation and Infiltration to be the most significant predictor of the two energy targets, then outside dry-bulb temperature - results with clear practical consequences on ventilation management strategies of educational buildings in hot climates.

There were six regression algorithms under supervision which were compared. The ensemble-based techniques namely Gradient Boosting Machines and XGBoost showed unquestionable dominance over the singular model techniques in both target variables. The maximum predictive accuracy of Gradient Boosting was found on cooling energy (MAE: 6.420 kWh, R 2: 0.990) and equipment energy (MAE: 0.2716 kWh, R 2: 0.9997), which validated the usefulness of sequential learning of ensembles to capture the detailed nonlinear interactions of thermal, solar, and The results of the Support Vector Regression demonstrated critically poor performance on both targets without any significant tuning of the kernel, which highlights the significance of algorithm choice in high-dimensional, seasonally varying building energy data.

The resulting prediction model forms a solid data-driven basis to the subsequent phase of intelligent building control - where precise daily energy forecasting is inputted into nature-inspired optimisation algorithms to automatically produce smart HVAC scheduling choices that optimize energy usage but maintain the occupant thermal comfort. Future research will build on this framework to include multi-zone disaggregated energy prediction, use real-time

https://doi.org/10.71058/jodac.v10i0714

occupancy sensing as a explicit model input, and test the transferability of trained models to climatically similar education buildings in Rajasthan. The open-access data stored on IEEE DataPort also adds to the reproducibility and benchmarking of the smart building energy research community.

References

[1] Bourhnane, S., Abid, M. R., Lghoul, R., Zine-Dine, K., Elkamoun, N., & Benhaddou, D. (2020). Machine learning for energy consumption prediction and scheduling in smart buildings. SN Applied Sciences, 2(2). https://doi.org/10.1007/s42452-020-2024-9

[2] Broujeny, R. S., Madani, K., Chebira, A., Amarger, V., & Hurtard, L. (2020). Data-driven living spaces' heating dynamics modeling in smart buildings using machine learning-based identification. Sensors (Switzerland), 20(4). https://doi.org/10.3390/s20041071

[3] Sari, M., Ali Berawi, M., Zagloel, T. Y., Madyaningarum, N., Miraj, P., Pranoto, A. R., Susantono, B., & Woodhead, R. (2023). MACHINE LEARNING-BASED ENERGY USE PREDICTION FOR THE SMART BUILDING ENERGY MANAGEMENT SYSTEM. Journal of Information Technology in Construction, 28, 622–645. https://doi.org/10.36680/j.itcon.2023.033

[4] Shah, A. S., Nasir, H., Fayaz, M., Lajis, A., Ullah, I., & Shah, A. (2020). Dynamic user preference parameters selection and energy consumption optimization for smart homes using deep extreme learning machine and bat algorithm. IEEE Access, 8, 204744–204762. https://doi.org/10.1109/ACCESS.2020.3037081

[5] Chaganti, R., Rustam, F., Daghriri, T., Díez, I. de la T., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Building Heating and Cooling Load Prediction Using Ensemble Machine Learning Model. Sensors, 22(19). https://doi.org/10.3390/s22197692

[6] Bekdaş, G., Aydın, Y., Isıkdağ, Ü., Sadeghifam, A. N., Kim, S., & Geem, Z. W. (2023). Prediction of Cooling Load of Tropical Buildings with Machine Learning. Sustainability (Switzerland), 15(11). https://doi.org/10.3390/su15119061

[7] AlZaabi, H., Shaalan, K., Ghazal, T. M., Khan, M. A., Abbas, S., Mago, B., Tomh, M. A. A., & Ahmad, M. (2023). Intelligent Energy Consumption For Smart Homes Using Fused Machine-Learning Technique. Computers, Materials and Continua, 74(1), 2261–2278. https://doi.org/10.32604/cmc.2023.031834

[8] Anjomshoaa, A., & Curry, E. (2021). Transfer Learning in Smart Environments †. Machine Learning and Knowledge Extraction, 3(2), 318–332. https://doi.org/10.3390/make3020016

[9] Mazhar, T., Malik, M. A., Haq, I., Rozeela, I., Ullah, I., Khan, M. A., Adhikari, D., Ben Othman, M. T., & Hamam, H. (2022). The Role of ML, AI and 5G Technology in Smart Energy and Smart Building Management. Electronics (Switzerland), 11(23). https://doi.org/10.3390/electronics11233960

[10] Afonso, J. A., Monteiro, V., & Afonso, J. L. (2023). Internet of Things Systems and Applications for Smart Buildings. In Energies (Vol. 16, Issue 6). MDPI. https://doi.org/10.3390/en16062757

https://doi.org/10.71058/jodac.v10i0714

[11] Elsisi, M., Tran, M. Q., Mahmoud, K., Lehtonen, M., & Darwish, M. M. F. (2021). Deep learning-based industry 4.0 and internet of things towards effective energy management for smart buildings. Sensors (Switzerland), 21(4), 1–19. https://doi.org/10.3390/s21041038

[12] Mariano-Hernández, D., Hernández-Callejo, L., Solís, M., Zorita-Lamadrid, A., Duque-Perez, O., Gonzalez-Morales, L., & Santos-García, F. (2021). A data-driven forecasting strategy to predict continuous hourly energy demand in smart buildings. Applied Sciences (Switzerland), 11(17). https://doi.org/10.3390/app11177886

[13] Ghasemkhani, B., Yilmaz, R., Birant, D., & Kut, R. A. (2022). Machine Learning Models for the Prediction of Energy Consumption Based on Cooling and Heating Loads in Internet-of-Things-Based Smart Buildings. Symmetry, 14(8). https://doi.org/10.3390/sym14081553

[14] Maghraoui, A. El, Ledmaoui, Y., Laayati, O., Hadraoui, H. El, & Chebak, A. (2022). Smart Energy Management: A Comparative Study of Energy Consumption Forecasting Algorithms for an Experimental Open-Pit Mine. Energies, 15(13). https://doi.org/10.3390/en15134569

[15] Afzalan, M., & Jazizadeh, F. (2020). A Machine Learning Framework to Infer Time-of-Use of Flexible Loads: Resident Behavior Learning for Demand Response. IEEE Access, 8, 111718–111730. https://doi.org/10.1109/ACCESS.2020.3002155

[16] Kharbouch, A., Berouine, A., Elkhoukhi, H., Berrabah, S., Bakhouya, M., El Ouadghiri, D., & Gaber, J. (2022). Internet-of-Things Based Hardware-in-the-Loop Framework for Model-Predictive-Control of Smart Building Ventilation. Sensors, 22(20). https://doi.org/10.3390/s22207978

[17] Marinakis, V. (2020). Big data for energy management and energy-efficient buildings. Energies, 13(7). https://doi.org/10.3390/en13071555

[18] Balaji, S., & Karthik, S. (2023). Energy Prediction in IoT Systems Using Machine Learning Models. Computers, Materials and Continua, 75(1), 443–459. https://doi.org/10.32604/cmc.2023.035275