FMDB *Transactions on Sustainable Energy Sequence*

Data-Driven and Domain Adaption Framework for Optimizing Energy Usage Forecasting in Smart Buildings Using ARIMA Model

D. Suraj[1,*], Najib Sulthan[2], R. Balaji[3], R. Swathi[4], Charlotte Roberts[5], B. Vaidianathan[6]

[1,2,3,4]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
[5]Australian Graduate School of Engineering (AGSE), University of New South Wales, Sydney, Australia.
[6]Department of Electronics and Communication Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
surajdoraiswamy@gmail.com[1], najib4cr7@gmail.com[2], balaji200219@gmail.com[3], swathir2@srmist.edu.in[4], charlotteroberts.gs@gmail.com[5], vaidianathan@dhaanishcollege.in[6]

Abstract: Building energy demand and energy system supply are increasingly balanced by energy storage and short-term DSM. This is necessary due to fluctuating renewable energy supplies and rising building power use. Worldwide, structures utilize tons of energy. Greenhouse gas emissions and operational expenses decrease with construction energy efficiency. Forecasting and optimization algorithms can solve challenges, including supply chains (inventory optimization), traffic, and sustainable energy system battery/load/production scheduling for carbon-free energy generation. We often solve optimization problems that need forecasting due to uncertain future values. Predicting and optimizing are challenging; therefore, little research has been done. Our method uses building energy modeling professionals' data to forecast neighboring building types for new or unknown building types. After training, we utilize the models to estimate energy usage for the k-closest building types and combine the predictions using a weighted average. We used time-series decomposition to detect uncertainty and a hybrid model to close this gap: The concepts encode static features and predictable patterns in time-series simulation results. The model learns latent performance differences and calibrates output using outcomes and history records. Historical data predicts public building energy ARIMA use. Our method covers data processing, training, validation, and forecasting. We measured our method. Mixed integer linear optimization and ARIMA projected most accurately.

Keywords: Gradient Boosting (GB); Combined Cycle Power Plant; Auto-Regressive Integrated Moving Average (ARIMA); Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX); Long Short-Term Memory.

Received on: 29/10/2023, **Revised on:** 02/01/2024, **Accepted on:** 03/03/2024, **Published on:** 09/06/2024

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSES

DOI: https://doi.org/10.69888/FTSES.2024.000187

Cite as: D. Suraj, N. Sulthan, R. Balaji, R. Swathi, C. Roberts, and B. Vaidianathan, "Data-Driven and Domain Adaption Framework for Optimizing Energy Usage Forecasting in Smart Buildings Using ARIMA Model," *FMDB Transactions on Sustainable Energy Sequence*., vol. 2, no. 1, pp. 1–11, 2024.

1. Introduction

Many complex real-world procedures are based on optimization problems that must be solved over an undetermined future. For instance, staffing rosters and supply chain inventories must be scheduled based on projections of future customer demand. This optimization will also be essential to the worldwide effort to cut CO emissions. Renewable energy production is

[*]Corresponding author.

characterized by fluctuation across time and difficulty in easily adjusting output in response to demand. Because future production and demand are unpredictable, demand must be planned to maximize supply where it can, and energy storage devices like batteries must be scheduled to make up for any shortage [1]. Typically, these issues are resolved by projecting the future and using that as the real input for the optimization process. While this is quick, it doesn't consider how unclear the forecast is. Using resilient optimization or stochastic optimization with probabilistic forecasts as inputs rather than point forecasts is one technique to deal with uncertainty [2].

Electrical energy is a necessary component of all industrial and production systems for a single household. Power plants must efficiently produce and distribute energy based on customer demand to make the most use of the natural resources that are currently available [3]. These days, energy usage data from multiple sources may be tracked and recorded by smart home devices throughout the house. Energy firms may find this extremely detailed data useful in efficiently managing the production and distribution of power [4].

The ecosystem's fundamental building block is its inhabitants. Occupants must control the building's systems to create the appropriate environment because they use the structure's amenities [5]. The building has systems and equipment to regulate and uphold the ideal atmosphere for the residents and diagnostic systems to guarantee reliable functioning. Energy is needed for building operations, and electricity is the main energy source. Buildings receive their electrical energy through a power distribution system, and with the development of smart grids, buildings can communicate and trade excess energy with one another and the energy supplier [6].

Alongside strategies like incentive design and price modification, researchers and industry leaders have tried to apply intelligent sensing control and automation systems to regulate energy usage and increase occupant comfort more efficiently [7]. A system that can sufficiently account for the heterogeneity of user-to-device and device-to-device interactions in buildings is required due to the proliferation of internet-of-things (IoT) devices [8]. It has also been essential to conclude that there is sparse data in certain cases and plenty of data in others. Driven by the abovementioned obstacles, machine learning (ML) has been used as a technology in smart buildings with progressively significant ramifications [9]. Machine learning (ML) algorithms process data, derive valuable insights from it, and apply it to activities like forecasting, prediction, and control. This article examines the current state of machine learning applications in smart building systems. We discuss current conventional approaches and how ML-based solutions frequently equal or surpass them [10].

The proliferation of data and growing processing power in recent years have resulted in a notable boost in the performance of data-driven prediction models. Data-driven techniques, such as deep neural networks (DNN), decision trees, support vector machines (SVM), and other machine learning techniques, are used in many current energy consumption modeling studies [11]. The data-driven model is trained using centralized data in these earlier approaches. Due to the inherent difficulties of predicting and optimization, the combined complexity of Predict+Optimize models in the research may not translate to practical issues. It could be the case that the combined problem has too many computation steps or that the problem instances must be described [12]. Complex optimization and a real-world data set are needed to solve realistic challenges. We observe that there aren't many issues in methodically ascertaining this study area's state of the art. Standard benchmark problems can be established through competitions [13].

As far as we know, there is only one competition in this field: the ICON Challenge on Forecasting and Scheduling hosts. In order to schedule server activities in a way that minimizes energy costs, it was necessary to forecast a single time series, the energy price [14]. This challenge leaned significantly toward optimization with a reasonably straightforward prediction task and a challenging optimization problem. The competition winner used heuristics to produce a preliminary solution, which a hill climbing algorithm then refined [15].

Analyzing a set of data points indexed in time order is known as time series analysis. The goal of the analysis is to gather and examine historical data from a time series to create a suitable model that captures the fundamental structure of the data. After that, forecasts—or future values for the series are produced using this model [16]. The dissipative dynamics of excitation energy transfer (EET) in systems resembling the photosynthetic open quantum system regime are the subject of the data analysis in this work [17].

Sometimes, information on the underlying dynamical correlations can be encoded early in the evolution of open quantum systems. Consequently, understanding the short-term development of open quantum systems might help us derive their long-term dynamics [18]. This conjecture makes it possible to avoid using direct long-term simulations. It is desirable to develop a method that can accurately predict long-time dynamics of open quantum systems and, to some extent, eliminate the need for direct calculations [19]. This is because the simulation of numerically exact methods to describe the dynamics of open quantum systems often requires enormous computational resources that scale exponentially with the size of the system under study [20].

A crucial component of intelligent buildings is to offer the best possible living environments while adhering to various energy and regulatory requirements. Energy consumption management is required to have the best cost of comfort and peace of mind in the structures [21]. As a result, the building has several intelligent features and appliances installed to control the indoor environment. Most energy used in buildings provides thermal and refrigeration comfort, such as water supply facilities, sanitary spas, lighting-related amenities, and heating and cooling systems. Additionally, depending on the type of building, different equipment is installed in each one, and each piece of equipment uses energy [22]. Energy is, therefore, used differently in each structure to meet the needs of its occupants.

Approximately 40% of all energy usage is attributed to buildings. One of the most crucial elements of smart cities is building energy management. Urban development's sustainability index is a social function of each developing city's local energy consumption and production [23]. Each developing city's energy generation and direct consumption determine the sustainability index of efficient urban development. A significant portion of the energy in an urban area is attributed to construction energy consumption. Accurately estimating and forecasting the production and use of energy in the building sector is the goal of multiple methodologies [24]. Overall, two different necessary steps can be useful at the building level in this direction. Several models grounded on physical principles are developed to provide a mathematical justification for thermal dynamics and energy behavior. These fundamental models are classified according to the kind of building and useful parameters [25]. Other sorts of models that are used to estimate energy consumption based on variables affecting climate and energy costs are statistical models. This viewpoint shows demand and consumption forecasting is crucial to creating smart cities[26].

As a branch of artificial intelligence (AI), machine learning (ML) based methods can offer a useful platform for modeling by considering various aspects. Lately, ML-based methods have substantially contributed to implementing trustworthy estimate models [27]. Several studies have used machine learning (ML) approaches in various disciplines to estimate the thermal conductivity of water-alumina nanofluids. These techniques include multi-layered perceptrons (MLP), radial basis functions (RBF), cascade feedforward (CFF), and generalized regression neural networks (GRNN) [28].

2. Existing System

Hydroelectricity has been widely used worldwide; in Canada, Norway, and Brazil, it continues to be the primary source of electricity. Even with the current diversification, about 65% of Brazil's electricity is produced from hydroelectricity. It is crucial to meet the national and international targets for lowering carbon emissions because it is the biggest non-polluting source in the nation. The resilience of the hydropower sector to global climate change is challenged by the need to meet energy and environmental sustainability standards. These changes impact power production by modifying seasonality, increasing streamflow unpredictability, and increasing reservoir evaporation losses. As a result, the architects of the current system determined which of the 27 climate indices were most important for enhancing the performance of the models used in the monthly seasonal streamflow series forecasting. Three machine learning models (support vector regression, extreme learning machine, and kernel ridge regression) and one linear model (seasonal autoregressive integrated moving average with exogenous factors, or SARIMAX) used a NOAA Physical Sciences Laboratory database as exogenous variables. The feature-selection method was a random forest with recursive feature elimination [29].

The outcomes made it possible to determine which set of indices was most pertinent for the plants under study, which enhanced forecasts of streamflow. In the watersheds of southern Brazil's primary hydro basins, this study examined the impact of 27 climatic indices on streamflow forecasts for a group of hydroelectric units. The effects of the ENSO in Northeastern and North America typically have the reverse effect in the southern regions. In northern and northeastern Brazil, a positive ENSO phase decreases precipitation, whereas a negative phase increases streamflow and precipitation.

Additionally, the results demonstrated a correlation between the ENSO_PI and MEI_v2 indices and stations in Brazil's northeast (NE) and north (N) in CI2 and CI3. Climate Indices Impact on Monthly Streamflow Series Forecasting, J. F. D. Toledo et al. showed the best link (CI1) with stations in the northeast, north, and southeast (SE). A random forest with recursive feature reduction was used to pick the most relevant climatic indicators. These were then identified using three machine learning models (SVR, KRR, and ELM) and one linear model (SARIMAX). Furthermore, the creators of the current approach employed techniques to categorize the climatic indices according to how crucial it was to lower the average absolute inaccuracy of the forecasts. The results showed gains when models containing climate indices in the input data were used as the exogenous variables.

Consideration of the CI improved the resolution of the case studies, particularly those along the equatorial line in the northeast and north-Brazilian regions. The case studies pertain to different river bases. These results are significant because precise streamflow prediction is critical to pricing strategies and operations planning. It is advised that alternative approaches, such as filters, be used for variable selection and that a historical series of climate indicators with time lags be tested to maintain the ongoing work on this paper. A study analyzing the methods for handling multi-step horizons must be established for forecasting

models. Additionally, among the linear, nonlinear, and combination approaches that can be employed in light of this work's viewpoints, a few models have emerged in recent times.

3. Proposed System

In actuality, an appropriate model is fitted to a particular time series, and using the values of the available data, the underlying model's corresponding parameters are calculated. Time Series Analysis is fitting a time series to an appropriate model. It includes techniques aimed at comprehending the series' nature and is frequently helpful for predicting and simulating future events. Historical observations are gathered and examined to create an appropriate mathematical model that accurately depicts data creation for the series. The model is then used in time series forecasting to anticipate future events.

We classified the uncertainty within the building performance gap between predictions and historical data. Additionally, our study showed that data-driven techniques and knowledge-based approaches play complementary roles in minimizing uncertainty. We suggested a technique to enable the integration of data-driven models and knowledge-based approaches into a hybrid framework. The framework significantly improves accuracy by effectively utilizing data from dynamic historical records and static characteristics while requiring fewer specific construction details. This framework was implemented in the domain to close the performance disparity.

The core concept of our approach involves synthesizing expert knowledge of building behavior with advanced machine-learning techniques. By incorporating domain expertise into the modeling process, our framework can effectively capture nuanced relationships and dependencies within the building system that might be challenging to discern solely from historical data. At the heart of our methodology lies a multi-step process:

- Knowledge Elicitation and Representation: We begin by eliciting and formalizing domain knowledge from building experts. This knowledge encompasses structural attributes, material properties, HVAC systems, and other factors influencing building performance.
- Feature Engineering and Data Preprocessing: Static and dynamic features are extracted from the collected data. Static features encompass architectural attributes, geographical location, and building specifications. Dynamic features include time-series data related to energy consumption, occupancy patterns, and environmental conditions.
- Hybrid Model Development: We develop a hybrid modeling framework integrating knowledge-based rules with data-driven algorithms. Knowledge-based rules provide interpretability and enforce constraints derived from expert insights. At the same time, data-driven algorithms, such as recurrent neural networks or gradient-boosted trees, capture complex temporal patterns in the data.
- Model Training and Validation: The hybrid model is trained on historical data and validated using rigorous testing procedures. We assess the model's performance against traditional data-driven approaches and baseline methods to quantify the improvement achieved by incorporating domain knowledge.
- Performance Evaluation and Comparison: We evaluate our hybrid framework's accuracy, robustness, and generalizability against existing techniques. Comparative analysis sheds light on the scenarios where the hybrid approach excels and offers significant advantages over purely data-driven or rule-based methods.

4. Methodology

The methodology involves a comprehensive approach to address optimization challenges in uncertain future scenarios. It includes developing and implementing robust optimization and stochastic optimization techniques to effectively tackle supply chain management and renewable energy production forecast uncertainty. The methodology also includes using machine learning (ML) algorithms to analyze data gathered from Internet-of-Things (IoT) devices installed in smart buildings, such as regression neural networks and deep neural networks.

Furthermore, the methodology uses ML-based techniques to build predictive energy production and consumption models. These models are crucial for optimizing energy usage, improving occupant comfort, and enhancing overall building performance. The methodology also involves leveraging ML algorithms' insights to develop innovative energy optimization and building automation solutions.

Moreover, the methodology emphasizes the importance of accurate prediction models for supporting the development of smart cities and sustainable urban development initiatives. By integrating optimized solutions and ML algorithms into existing building infrastructure, the methodology aims to promote energy-efficient practices and technologies, ultimately contributing to sustainability and environmental conservation efforts.

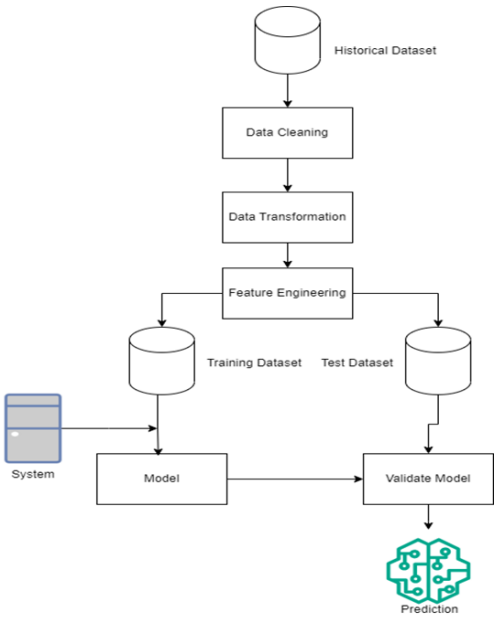


Figure 1: Architecture Diagram

Figure 1 represents the architecture diagram. At the center of the diagram is the core component, which represents the machine learning model responsible for energy demand prediction. Surrounding this core component are several key elements: Data Sources represent the various data sources used to train and validate the machine learning model. This may include historical energy consumption data, weather data, building characteristics, and other relevant information.

Data Preprocessing Module: This module is responsible for cleaning and preprocessing the raw data before it is fed into the machine learning model. It handles tasks such as data normalization, missing values, and feature engineering.

Training Module: This module encompasses the training processes for the machine learning model. It includes tasks such as model initialization, optimization, and evaluation using training data.

Validation Module: This module validates the trained model using separate validation datasets. It assesses the model's performance and generalization ability by comparing its predictions with actual energy consumption data.

Prediction Module: Once the model is trained and validated, the prediction module utilizes it to make energy demand forecasts. It takes input data from real-time or historical sources and generates predictions for future energy consumption.

Feedback Loop: This component represents the feedback loop between the prediction and training modules. It allows for continuous model improvement by incorporating new data and updating model parameters based on prediction errors. Overall, the architecture diagram illustrates how data flows through different modules within the system, from data acquisition to prediction generation, facilitating the efficient operation of the energy demand prediction system.

5. Module Description

The entire process is divided into three modules.

5.1. Module 1: Data preprocessing

A time series is a collection of observations taken regularly and recorded in an even interval. Although time series data is often opaque, it contains much information. Unordered timestamps, missing values (or timestamps), outliers, and noise in the data are common issues with time series. Managing the missing values is the most challenging of all the issues discussed. Most Time Series data is found in unstructured formats, meaning timestamps may be mixed up or arranged incorrectly.

Additionally, the date-time column typically contains a string data type by default. Therefore, before performing any operations on it, it is imperative to convert the data-time column to a datetime datatype. It might be difficult to handle missing numbers in

time series data. Because the order in which values are received is important, traditional imputation approaches are unsuitable for time-series data (Figure 2).

	A	B	C	D	E	F	G	H	I
1	LCLid	day	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
2	MAC000131	15-12-2011	0.485	0.432045455	0.868	22	0.239145797	9.505	0.072
3	MAC000131	16-12-2011	0.1415	0.296166669	1.1160001	48	0.281471318	14.2160001	0.031
4	MAC000131	17-12-2011	0.1015	0.1898125	0.685	48	0.188404686	9.111	0.064
5	MAC000131	18-12-2011	0.114	0.218979167	0.676	48	0.202919279	10.511	0.065
6	MAC000131	19-12-2011	0.191	0.325979167	0.788	48	0.259204962	15.647	0.066
7	MAC000131	20-12-2011	0.218	0.3575	1.077	48	0.28759657	17.16	0.066
8	MAC000131	21-12-2011	0.1305	0.235083333	0.705	48	0.222069649	11.284	0.066
9	MAC000131	22-12-2011	0.089	0.221354167	1.094	48	0.267238875	10.625	0.062
10	MAC000131	23-12-2011	0.1605	0.291125	0.749	48	0.249076048	13.974	0.065
11	MAC000131	24-12-2011	0.107	0.169	0.613	47	0.150684669	7.943	0.065
12	MAC000131	25-12-2011	0.2175	0.3391875	0.866	48	0.263101199	16.281	0.069
13	MAC000131	26-12-2011	0.1495	0.261708333	0.838	48	0.244792744	12.562	0.066
14	MAC000131	27-12-2011	0.143	0.274	0.778	48	0.252127458	13.152	0.068
15	MAC000131	28-12-2011	0.1455	0.300520833	1.207	48	0.298680288	14.425	0.066
16	MAC000131	29-12-2011	0.152	0.307041667	0.888	48	0.264454634	14.738	0.066
17	MAC000131	30-12-2011	0.135	0.276854167	0.782	48	0.261185757	13.289	0.064
18	MAC000131	31-12-2011	0.1515	0.325729167	1.252	48	0.309888294	15.635	0.066
19	MAC000131	01-01-2012	0.151	0.256020833	0.812	48	0.225249412	12.289	0.068
20	MAC000131	02-01-2012	0.134	0.252083333	0.851	48	0.23721297	12.1	0.068
21	MAC000131	03-01-2012	0.1475	0.2355	0.674	48	0.209995339	11.304	0.068
22	MAC000131	04-01-2012	0.101	0.216270833	0.731	48	0.215205764	10.381	0.065
23	MAC000131	05-01-2012	0.146	0.331020833	0.786	48	0.27337337	15.889	0.063
24	MAC000131	06-01-2012	0.1025	0.223645833	0.765	48	0.241187002	10.735	0.064
25	MAC000131	07-01-2012	0.0995	0.137395833	0.572	48	0.094495466	6.595	0.065
26	MAC000131	08-01-2012	0.137	0.2305	0.857	48	0.221244449	11.064	0.068
27	MAC000131	09-01-2012	0.1345	0.260354167	0.771	48	0.243380491	12.497	0.065
28	MAC000131	10-01-2012	0.1435	0.201291667	0.694	48	0.16980514	9.662	0.068

Figure 2: Raw Data

The following steps are involved in the data processing (Figure 3):

- The data set is processed; the timestamps are stored in an array, and the energy consumption measurements are kept in a matrix.
- Date and time objects are created from the timestamps.
- Two new matrices are constructed: one holds the multidimensional input feature vectors from the neural network, and the other comprises energy consumption measures to assess errors.

```
energy = daily_df.groupby('day')[['energy_sum']].sum()
```

```
energy = energy.merge(house_count, on = ['day'])
```

```
energy = energy.reset_index()
```

```
energy.head()
```

	day	energy_sum	LCLid
0	2011-11-23	90.38500	13
1	2011-11-24	213.41200	25
2	2011-11-25	303.99300	32
3	2011-11-26	420.97600	41
4	2011-11-27	444.88300	41

```
energy["energy_per_household"] = energy["energy_sum"] / energy["LCLid"]
```

```
energy["day"] = pd.to_datetime(energy["day"])
```

```
energy.day = pd.to_datetime(energy.day,format='%Y-%m-%d').dt.date
```

```
energy
```

	day	energy_sum	LCLid	energy_per_household
0	2011-11-23	90.38500	13	6.95269
1	2011-11-24	213.41200	25	8.53648
2	2011-11-25	303.99300	32	9.49978
3	2011-11-26	420.97600	41	10.26771

Figure 3: Data Preprocessing

5.2. Module 2: Feature Processing

The measurement vectors used in error evaluation have the same features as the input feature vectors. We experimented with several attributes for these vectors during our experimental evaluation to get the best accurate findings. In our first experiment, we divide the timestamp into four numbers: year, month, day, and half hour. At this point, the input data is represented by an array of four-dimensional vectors.

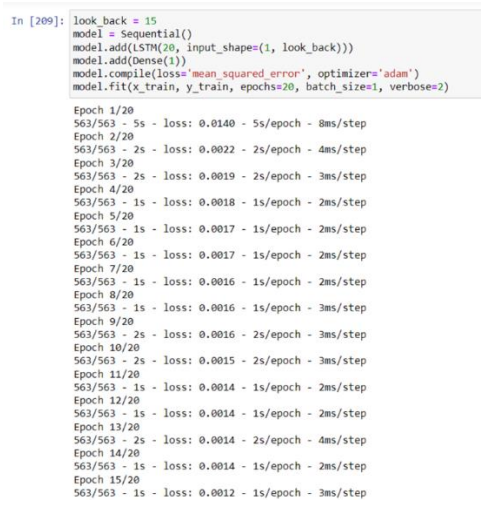
We added as much seasonal data as we could to the input vectors. We used the day of the week, day of the year, and week of the year features in addition to the ones from the previous phase. In our third experiment, we merged a regression-based strategy with a time series forecasting method to significantly increase the prediction accuracy. The plan will incorporate measurements from prior half-hours as input features via a sliding window approach. One drawback to this method is that the network can only forecast one half-hour's consumption at a time; consequently, the following half-hour's consumption must be provided as an input feature. It is an extremely accurate prediction for a short forecasting horizon. On the other hand, the mistake propagates for a big horizon as we move farther away from the training interval's end.

5.3. Module 3: Time Series Data Training and Prediction

The data set is split into two subsets—one for training and one for validation for the training and validation cycle. We train and validate the network over a predetermined number of epochs. The following procedures are part of a training phase. The training subset is then divided into batches, and each batch goes through the network through a forward and a backward propagation pass. The batch is fed into the network during the forward pass, and the output results represent the anticipated consumption levels. The Mean Absolute Error is obtained by comparing the actual readings with the projected usage. The Adam optimization approach modifies the network's weights during the backpropagation stage to reduce mistakes. The trained neural network model is then saved after the error on the training period is calculated by averaging the errors for each batch.

The data subset used for validation must begin at the end of the training subset since we solve the energy forecasting problem as a time series forecasting problem using a sliding window method. Therefore, several measurements from the training subset will be used as input values for several validation examples, depending on the size of the sliding window. Every input example is processed one at a time in the validation step (the examples are not divided into batches). The predicted value obtained from feeding the validation example into the network is then normalized and added to the input data of the subsequent input example. Subsequently, the Mean Absolute Percentage Error and the Mean Absolute Error are calculated. Lastly, for the two error metrics stated above, the total error value over the entire validation subset is calculated.

During the testing phase, the energy consumption on the testing data set is predicted using a pre-trained model. Predictions are based on the parameters, weights, and the network's state following a specific epoch. Forecasting is possible on a specified horizon, with the first validation example following the last training example serving as the horizon's beginning point. This functionality is quite helpful to have. A portion of the validation subset that aligns with the selected forecasting horizon is usable. With the neural network model that has been trained, we may select several horizons and observe how the error varies with horizon length (Figure 4).



```
In [209]: look_back = 15
          model = Sequential()
          model.add(LSTM(20, input_shape=(1, look_back)))
          model.add(Dense(1))
          model.compile(loss='mean_squared_error', optimizer='adam')
          model.fit(x_train, y_train, epochs=20, batch_size=1, verbose=2)

Epoch 1/20
563/563 - 5s - loss: 0.0140 - 5s/epoch - 8ms/step
Epoch 2/20
563/563 - 2s - loss: 0.0022 - 2s/epoch - 4ms/step
Epoch 3/20
563/563 - 2s - loss: 0.0019 - 2s/epoch - 3ms/step
Epoch 4/20
563/563 - 1s - loss: 0.0018 - 1s/epoch - 2ms/step
Epoch 5/20
563/563 - 1s - loss: 0.0017 - 1s/epoch - 2ms/step
Epoch 6/20
563/563 - 1s - loss: 0.0017 - 1s/epoch - 2ms/step
Epoch 7/20
563/563 - 1s - loss: 0.0016 - 1s/epoch - 2ms/step
Epoch 8/20
563/563 - 1s - loss: 0.0016 - 1s/epoch - 3ms/step
Epoch 9/20
563/563 - 2s - loss: 0.0016 - 2s/epoch - 3ms/step
Epoch 10/20
563/563 - 2s - loss: 0.0015 - 2s/epoch - 3ms/step
Epoch 11/20
563/563 - 1s - loss: 0.0014 - 1s/epoch - 2ms/step
Epoch 12/20
563/563 - 1s - loss: 0.0014 - 1s/epoch - 2ms/step
Epoch 13/20
563/563 - 2s - loss: 0.0014 - 2s/epoch - 4ms/step
Epoch 14/20
563/563 - 1s - loss: 0.0014 - 1s/epoch - 2ms/step
Epoch 15/20
563/563 - 1s - loss: 0.0012 - 1s/epoch - 3ms/step
```

Figure 4: Training Module

5.4. Efficiency of this Model

The proposed forecasting system exhibits notable efficiency advantages, including reduced communication costs through clustering techniques, significantly faster performance than models trained without clustering, and suitability for time series prediction tasks. By leveraging advanced clustering algorithms and predictive modeling techniques, the system enhances forecasting accuracy and demonstrates good generalization ability despite limited sample sizes. Its design and methodology are tailored specifically for time series data analysis, ensuring reliable predictions for various applications. The system offers a comprehensive solution for efficient and accurate time series forecasting, making it suitable for diverse forecasting tasks across industries.

Out[253]:

day	energy_per_household	temperatureMax	holiday_id	prediction	diff
2014-02-27	10.35635	10.31000	0	11.54957	-1.19322
2014-02-26	10.20295	11.29000	0	11.33378	-1.13083
2014-02-25	10.29500	11.43000	0	11.36390	-1.06891
2014-02-21	10.51813	10.15000	0	11.50228	-0.98416
2014-02-19	10.67462	10.13000	0	11.40391	-0.72928
2014-02-18	10.78190	10.13000	0	11.40995	-0.62806
2014-02-24	10.41140	14.23000	0	11.02174	-0.61034
2014-02-20	10.57384	12.50000	0	11.17090	-0.59706
2014-02-03	11.28001	7.99000	0	11.86853	-0.58852
2014-02-04	11.09558	8.88000	0	11.65922	-0.56364
2014-02-13	11.28574	7.37000	0	11.83978	-0.55404
2014-02-22	10.77624	11.63000	0	11.23840	-0.46216
2014-02-07	10.97232	9.81000	0	11.41089	-0.43858
2014-02-17	10.97957	10.67000	0	11.29412	-0.31455
2014-02-10	11.26418	8.78000	0	11.54175	-0.27757
2014-02-11	11.45265	7.51000	0	11.70625	-0.25360
2014-01-30	11.68517	5.94000	0	11.81745	-0.13229
2014-02-15	11.49047	9.90000	0	11.61934	-0.12887
2014-02-08	11.56930	9.13000	0	11.54507	0.02423
2014-02-06	11.44540	9.81000	0	11.41876	0.02664
2014-02-12	11.67910	8.83000	0	11.64772	0.03138
2014-02-05	11.41510	9.64000	0	11.36846	0.04665
2014-02-16	11.58216	9.98000	0	11.52868	0.05348
2014-02-01	11.71058	9.72000	0	11.56516	0.14542
2014-01-31	11.85796	8.83000	0	11.62874	0.22922
2014-02-23	11.48041	11.94000	0	11.21545	0.26496
2014-02-09	12.20297	8.18000	0	11.73241	0.47056
2014-02-02	12.07816	9.30000	0	11.53833	0.53984
2014-02-14	11.81691	12.02000	0	11.23922	0.57770

Figure 5: Predicted Results

Figure 5 shows the predicted results, which the well-trained model does. The forecasting model generates predictions of future energy consumption based on the input time series data and other relevant features. These predictions can be numerical values representing expected energy usage for specific time intervals in the future.

Moreover, the model may also provide uncertainty estimates, confidence intervals, and predicted values. These uncertainty estimates help stakeholders make informed decisions based on the reliability of the forecasts.

In summary, the input to our paper includes time series energy consumption data and additional relevant features. At the same time, the output consists of predictions of future energy consumption and associated uncertainty estimates.

6. Discussions

The primary objective of this study was to develop an integrated framework for optimizing energy usage forecasting in smart buildings by leveraging both data-driven methodologies and domain-specific knowledge. The proposed framework aims to enhance energy consumption predictions' accuracy while addressing data variability and model adaptability challenges in dynamic building environments. In this paper, we employed Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) models as foundational components of our forecasting framework.

ARIMA models were utilized to capture the time-series dependencies and trends in energy consumption data, allowing for the prediction of future consumption patterns based on historical observations. SARIMAX models extended this capability by

incorporating exogenous variables such as weather conditions, occupancy patterns, and building characteristics. This significantly improved forecasting accuracy by accounting for external influences on energy usage.

Throughout the development of this framework, several challenges were encountered, primarily related to data preprocessing, model selection, and adaptation to dynamic building conditions:

Data Variability and Quality: The variability in sensor data quality and availability posed a significant challenge, requiring robust data preprocessing techniques to handle missing values, outliers, and inconsistencies.

Model Parameter Tuning: Fine-tuning the parameters of ARIMA and SARIMAX models, especially in the context of multivariate time series forecasting, proved complex and time-consuming. Optimizing model hyperparameters to achieve optimal forecasting performance was critical to our approach.

Domain Adaptation: Adapting generic forecasting models like ARIMA and SARIMAX to specific building contexts with diverse operational characteristics and environmental influences presented challenges. Incorporating domain knowledge effectively into the model architecture was essential for achieving accurate predictions.

One of the key outcomes of this paper is the achievement of enhanced forecasting accuracy in energy usage prediction for smart buildings. By integrating domain-specific knowledge with sophisticated data-driven models like ARIMA and SARIMAX, the framework demonstrated improved accuracy compared to traditional approaches. This heightened accuracy is crucial for optimizing energy consumption patterns, enabling more informed decision-making and resource allocation in building operations. The developed framework adapts to dynamic building conditions, essential for real-world applications in smart buildings. By incorporating exogenous variables such as weather conditions, occupancy patterns, and building characteristics into the forecasting models, the framework demonstrated robustness in handling external influences on energy usage. This adaptability ensures reliable and responsive energy consumption predictions, even in fluctuating operational environments.

The developed framework is designed for practical implementation and scalability across diverse building environments. Integrating data-driven methodologies with domain adaptation strategies offers a flexible and scalable solution for energy forecasting in smart buildings. The framework's modular architecture allows for customization and adaptation to specific building contexts, making it suitable for deployment in various smart building applications. One real-time use case for the proposed forecasting system could be in the energy sector, specifically for electricity demand forecasting. Energy companies could implement the system to predict short-term and long-term electricity demand based on historical consumption data, weather patterns, and other relevant factors. By accurately forecasting demand, energy providers can optimize resource allocation, ensure grid stability, and efficiently manage electricity generation. This would lead to cost savings, improved operational performance, and enhanced reliability in meeting customer demand. Additionally, the system's ability to adapt to changing usage patterns and incorporate real-time data updates would make it invaluable for energy companies seeking to navigate dynamic market conditions and regulatory requirements.

7. Conclusion

In this study, we investigated the development of a deep learning-based system for predicting energy demand, utilizing clustering and federated learning techniques for training. Federated learning enables the utilization of distributed client data by training local models without transmitting data. However, convergence in federated training can be slow due to differing data distributions across clients. To expedite convergence, clients with similar attributes can be clustered together, allowing for aggregation of model updates from clients within the same clusters. The outcomes of this research have potential applications in optimizing electricity grid operations and managing energy consumption in public buildings more efficiently.

Furthermore, energy consumption predictions could offer valuable insights into energy conservation strategies. The study's conclusions greatly impact how well the electrical grid functions and how public buildings manage energy use. Stakeholders can decide on resource allocation, peak load management, and energy conservation measures by employing deep learning models trained via federated learning to estimate energy demand accurately. Grid operators and building managers can optimize energy consumption patterns, save operating costs, and improve overall system efficiency with the help of these predictive capabilities.

Furthermore, our system's applicability goes beyond operational effectiveness to consider environmental sustainability. Our model's predictive energy consumption insights can guide the development and execution of focused energy conservation programs, promoting a sustainable culture in public infrastructure and urban planning. We open the door for more intelligent, environmentally friendly energy management strategies that support international efforts to achieve carbon neutrality and climate resilience by utilizing cutting-edge machine learning technology.

7.1. Future Enhancement

We envisage extending the capabilities of our federated learning technique by investigating new model designs and adaptive clustering algorithms to meet changing needs in diverse client contexts. Predictive analytics may also be extended to include multi-modal data sources and real-time feedback loops, which may present new possibilities for energy system optimization and the advancement of sustainable development. With the increasing prevalence of IoT technologies and the growing computational capabilities of machine learning (ML), there is a promising trajectory towards significantly improving occupant comfort and enhancing building energy performance. As ML algorithms continue to evolve and become more computationally feasible, they hold the potential to revolutionize the built environment by offering resilience and adaptability. In the future, ML could be pivotal in optimizing various aspects of building operations, including occupants' satisfaction levels, energy utilization efficiency, and cost-effectiveness. By leveraging the vast amount of data generated by IoT sensors and devices, ML algorithms can provide actionable insights and predictive analytics, enabling proactive decision-making and resource allocation to create more comfortable, sustainable, and energy-efficient buildings.

In order to further improve federated learning's effectiveness and flexibility in diverse client situations, adaptive clustering techniques may be explored in later studies. Adaptive clustering can maximize convergence rates and model collaboration by dynamically clustering clients according to changing contextual factors, including seasonal fluctuations or demographic shifts. Our system could continuously adjust to shifting data distributions if adaptive clustering techniques were implemented, guaranteeing reliable performance over various network topologies. Translating research results into practical applications requires investigating scalability and deployment issues. Future improvements should prioritize making our deep learning system scalable across various infrastructure configurations and large-scale networks. This entails looking at effective model compression methods, decentralized training approaches, and edge computing platform compatibility to enable smooth deployment in public infrastructure and smart grid scenarios.

Acknowledgment: I thank the respective institutions for their support and resources throughout this research. We sincerely thank SRM Institute of Science and Technology, Chennai, India; the Australian Graduate School of Engineering (AGSE) at the University of New South Wales, Sydney, Australia; and the Dhaanish Ahmed College of Engineering, Chennai, India.

Data Availability Statement: The data for this study can be made available upon request to the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines, obtaining informed consent from all participants.

References

1. A. Arjomandi-Nezhad, A. Ahmadi, S. Taheri, M. Fotuhi-Firuzabad, M. Moeini-Aghtaie, and M. Lehtonen, "Pandemic-aware day-ahead demand forecasting using ensemble learning," IEEE Access, vol. 10, no.1, pp. 7098–7106, 2022.
2. A. G. Usman et al., "Environmental modelling of CO concentration using AI-based approach supported with filters feature extraction: A direct and inverse chemometrics-based simulation," Sustain. Chem. Environ., vol. 2, no.6, p. 100011, 2023.
3. A. Gbadamosi et al., "New-generation machine learning models as prediction tools for modeling interfacial tension of hydrogen-brine system," Int. J. Hydrogen Energy, vol. 50, no.3, pp. 1326–1337, 2024.
4. B. D. Dimd, S. Voller, U. Cali, and O.-M. Midtgard, "A review of machine learning-based photovoltaic output power forecasting: Nordic context," IEEE Access, vol. 10, no.3, pp. 26404–26425, 2022.
5. B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach," IEEE Access, vol. 9, no.2, pp. 31191–31212, 2021.
6. B. Prasanth et al., "Maximizing Regenerative Braking Energy Harnessing in Electric Vehicles Using Machine Learning Techniques," Electronics, vol. 12, no. 5, p.12, 2023.
7. B. S. Alotaibi et al., "Sustainable Green Building Awareness: A Case Study of Kano Integrated with a Representative Comparison of Saudi Arabian Green Construction," Buildings, vol. 13, no. 9, p. 12, 2023.

8. D. Ganesh, S. M. S. Naveed, and M. K. Chakravarthi, “Design and implementation of robust controllers for an intelligent incubation Pisciculture system,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no. 1, pp. 101–108, 2016.
9. D. S. Kumar, A. S. Rao, N. M. Kumar, N. Jeebaratnam, M. K. Chakravarthi, and S. B. Latha, “A stochastic process of software fault detection and correction for business operations,” The Journal of High Technology Management Research, vol. 34, no. 2, p.11, 2023.
10. G. S. Sudheer, C. R. Prasad, M. K. Chakravarthi, and B. Bharath, “Vehicle Number Identification and Logging System Using Optical Character Recognition,” International Journal of Control Theory and Applications, vol. 9, no. 14, pp. 267–272, 2015.
11. H. M. Albert et al., “Crystal formation, structural, optical, and dielectric measurements of l-histidine hydrochloride hydrate (LHHCLH) crystals for optoelectronic applications,” J. Mater. Sci.: Mater. Electron., vol. 34, no. 30, p.17, 2023.
12. I. Abdulazeez, S. I. Abba, J. Usman, A. G. Usman, and I. H. Aljundi, “Recovery of Brine Resources Through Crown-Passivated Graphene, Silicene, and Boron Nitride Nanosheets Based on Machine-Learning Structural Predictions,” ACS Appl. Nano Mater., 2023, Press.
13. J. F. D. Toledo et al., “Climate indices impact in monthly streamflow series forecasting,” IEEE Access, vol. 11, no.9, pp. 21451–21464, 2023.
14. J. P. Singh, O. Alam, and A. Yassine, “Influence of geodemographic factors on electricity consumption and forecasting models,” IEEE Access, vol. 10, no.7, pp. 70456–70466, 2022.
15. M. A. Tripathi, K. Madhavi, V. S. P. Kandi, V. K. Nassa, B. Mallik, and M. K. Chakravarthi, “Machine learning models for evaluating the benefits of business intelligence systems,” The Journal of High Technology Management Research, vol. 34, no. 2, p.10, 2023.
16. M. A. Yassin et al., “Advancing SDGs : Predicting Future Shifts in Saudi Arabia ’ s Terrestrial Water Storage Using Multi-Step-Ahead Machine Learning Based on GRACE Data,” Saudi Arabia, 2024.
17. M. A. Yassin, A. G. Usman, S. I. Abba, D. U. Ozsahin, and I. H. Aljundi, “Intelligent learning algorithms integrated with feature engineering for sustainable groundwater salinization modelling: Eastern Province of Saudi Arabia,” Results Eng., vol. 20, no.7, p. 101434, 2023.
18. M. Chakravarthi and N. Venkatesan, “Design and implementation of adaptive model based gain scheduled controller for a real time non linear system in LabVIEW,” Research Journal of Applied Sciences, Engineering and Technology, vol. 10, no. 2, pp. 188–196, 2015.
19. M. Chakravarthi and N. Venkatesan, “Design and Implementation of Lab View Based Optimally Tuned PI Controller for A Real Time Non Linear Process,” Asian Journal of Scientific Research, vol. 8, no. 1, p.9, 2015.
20. M. Madhukumar, A. Sebastian, X. Liang, M. Jamil, and M. N. S. K. Shabbir, “Regression model-based short-term load forecasting for university campus load,” IEEE Access, vol. 10, no.1, pp. 8891–8905, 2022.
21. N. Mehboob, H. E. Z. Farag, and A. M. Sawas, “Energy consumption model for indoor cannabis cultivation facility,” IEEE Open J. Power Energy, vol. 7, no.6, pp. 222–233, 2020.
22. N. Sirisha, M. Gopikrishna, P. Ramadevi, R. Bokka, K. V. B. Ganesh, and M. K. Chakravarthi, “IoT-based data quality and data preprocessing of multinational corporations,” The Journal of High Technology Management Research, vol. 34, no. 2, p.13, 2023.
23. N. Venkatesan and M. K. Chakravarthi, “Adaptive type-2 fuzzy controller for nonlinear delay dominant MIMO systems: an experimental paradigm in LabVIEW,” Int. J. Adv. Intell. Paradig., vol. 10, no. 4, p. 354, 2018.
24. R. Jain et al., “Internet of Things-based smart vehicles design of bio-inspired algorithms using artificial intelligence charging system,” Nonlinear Eng., vol. 11, no. 1, pp. 582–589, 2022.
25. S. A. Nabavi, N. H. Motlagh, M. A. Zaidan, A. Aslani, and B. Zakeri, “Deep learning in energy modeling: Application in smart buildings with distributed energy generation,” IEEE Access, vol. 9, no.9, pp. 125439–125461, 2021.
26. S. Buragadda, K. S. Rani, S. V. Vasantha, and K. Chakravarthi, “HCUGAN: Hybrid cyclic UNET GAN for generating augmented synthetic images of chest X-ray images for multi classification of lung diseases,” Int. J. Eng. Trends Technol., vol. 70, no. 2, pp. 229–238, 2022.
27. S. I. Abba, A. G. Usman, and S. IŞIK, “Simulation for response surface in the HPLC optimization method development using artificial intelligence models: A data-driven approach,” Chemom. Intell. Lab. Syst., vol. 201, no. 4, p.15, 2020.
28. V. Vahidinasab, C. Ardalan, B. Mohammadi-Ivatloo, D. Giaouris, and S. L. Walker, “Active building as an energy system: Concept, challenges, and outlook,” IEEE Access, vol. 9, no.4, pp. 58009–58024, 2021.
29. Y. Tian, L. Sehovac, and K. Grolinger, “Similarity-based chained transfer learning for energy forecasting with big data,” IEEE Access, vol. 7, no.9, pp. 139895–139908, 2019.