Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 567~579
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp567-579

Evaluation of the impact of machine learning on the prediction of residential energy consumption

Richard Martín Machaca-Casani, Luis Alfredo Figueroa-Mayta, Joel Contreras-Nuñez
Facultad de Ingeniería, Escuela Académico Profesional de Ingeniería Eléctrica, Universidad Continental, Arequipa, Perú

Article Info

Article history:

Received Sep 1, 2025
Revised Oct 9, 2025
Accepted Oct 15, 2025

Keywords:

Applied artificial intelligence
Energy prediction
Machine learning
Predictive models
Residential consumption

ABSTRACT

The objective of this research was to compare the performance of machine learning models and traditional statistical methods for the prediction of residential energy consumption, using a dataset with relevant variables such as consumption, temperature, time of day, type of housing, and energy usage habits. A quantitative and comparative methodology was applied, involving data preprocessing, variable encoding, and normalization, as well as division into training and testing sets. The random forest, support vector machine (SVM), deep neural network (MLP), and linear regression models were trained and evaluated using standard metrics such as mean absolute error (MAE), root mean squared error (RMSE), and R^2 on test and cross-validation sets. Results show that SVM and linear regression achieved better accuracy and generalization capability, while random forest and the deep neural network exhibited lower explanatory power, reflected in negative R^2 values. Using the trained models, a projection of residential energy consumption for the 2026–2030 period was performed, revealing a generally increasing trend across all models, although with differences in the magnitude of the predictions. In conclusion, under the current conditions, traditional models demonstrate greater robustness, highlighting the need to tailor algorithm selection to the data context. These projections provide a valuable tool for future energy planning.

Corresponding Author:

Richard Martín Machaca-Casani
Facultad de Ingeniería, Escuela Académico Profesional de Ingeniería Eléctrica
Universidad Continental
Arequipa, Perú
Email: 73025874@continental.edu.pe

1. INTRODUCTION

Residential energy consumption represents a crucial challenge in the planning and operation of modern electrical systems. In the context of Arequipa, the electricity sector has experienced notable changes in its energy matrix. According to the Arequipa Chamber of Commerce [1], between January and September 2023, hydroelectric energy represented 63.9% of total production, while in 2024 it rose to 82.6%, accompanied by growth in solar generation. This variability highlights the need for predictive models capable of adapting to changes in generation and demand. Additionally, the Arequipa Electrical System accounts for 80.55% of the region's total consumption, mainly concentrated in urban areas [2], posing additional challenges for accurate forecasting.

This research begins with the following problem: how effective are machine learning models compared to traditional statistical methods in predicting residential energy consumption in Peruvian urban contexts? This question arises from the growing challenge of efficiently managing energy demand in residential settings, particularly in regions such as Arequipa, where intensive urbanization and consumption variability demand accurate and adaptable prediction tools. Previous studies have shown significant progress in the use of machine learning algorithms to estimate energy consumption in developed countries; however, there is a lack of empirical research in local contexts that directly contrasts these methods with traditional approaches such as linear regression or autoregressive integrated moving average (ARIMA) models. The central thesis argues that, although machine learning models can offer significant improvements in environments with complex and abundant data, their effectiveness in local contexts with limited data critically depends on the quality of the information and the precise tuning of the models.

The main theme of the research is the evaluation of the impact of machine learning models on the prediction of residential energy consumption, compared to traditional statistical methods. The proposed approach seeks to apply algorithms such as random forest, XGBoost, and neural networks to a set of historical data in order to compare their predictive accuracy against classical models such as ARIMA and linear regression. Metrics such as mean absolute error (MAE), root mean squared error (RMSE), and R^2 are used to identify which model offers the greatest accuracy and reliability. Furthermore, the practical applicability of these models in real scenarios of residential energy planning is analyzed, promoting their use to optimize consumption management in urban environments.

Locally and nationally, there is a significant absence of research integrating machine learning techniques to model residential energy consumption, with existing studies limited to traditional forecasting approaches based on polynomial functions or statistical projections [2], [3].

Machine learning (ML) is a subdiscipline of artificial intelligence that develops algorithms capable of learning patterns from data and making predictions or decisions without explicit programming [4]. Its importance has grown in sectors such as energy, where, according to [5], its dynamic adaptability improves consumption forecasting and sustainability. ML types include supervised learning, applied in energy prediction using variables such as temperature and consumption habits [6]; unsupervised learning, which identifies hidden patterns; and reinforcement learning, still emerging in the sector. Applications in electrical systems include demand forecasting [7], predictive maintenance [8], and early fault detection [6]. Despite its advances, ML faces challenges such as the need for large volumes of quality data, bias, and interpretability issues [9]. The recent recognition of Geoffrey Hinton with the Nobel Prize in Physics highlights its scientific impact [10]. In this context, ML significantly contributes to our research by offering more accurate tools to predict residential energy consumption, surpassing the limitations of traditional statistical methods and enabling more efficient and sustainable energy management.

Residential energy consumption forecasting is essential for the efficient planning of electrical systems, enabling the optimization of generation, distribution, and storage, as well as minimizing costs and emissions [11]. Factors such as climate, housing type, consumption habits, and dynamic electricity tariffs significantly influence consumption patterns [12]. Traditionally, statistical methods like multiple linear regression, ARIMA, and SARIMA have been used for this task [13], [14]; however, they have limitations in capturing nonlinear relationships and abrupt variations. Advances in machine learning have overcome these barriers using artificial neural networks, random forests, and support vector machines, achieving a significant reduction in prediction errors and greater adaptability [15]. Studies such as [16] demonstrated that hybrid models incorporating meteorological data can reduce RMSE by more than 15% compared to traditional models. Nevertheless, challenges such as data quality, generalization capacity, and model interpretability persist [17]. Thus, machine learning-based energy prediction is a key advancement toward smarter, more resilient, and sustainable electrical systems.

At the national level, Peru has shown growing interest in applying machine learning techniques to energy consumption prediction. A recent study proposed a second-degree polynomial function model to forecast the country's electricity consumption by 2030 [3], aiming to provide decision-makers with tools to face emerging challenges related to energy consumption. Furthermore, the Residential Energy Consumption and Usage Survey (ERCUE) 2019–2020 by OSINERGMIN [2] provides detailed data on residential consumption patterns, essential for developing predictive models that reflect regional realities. However, no indexed journal publications in the fields of electrical engineering or computer science were found nationally using ML models like neural networks, random forests, or support vector machines to compare their performance with traditional methods in the residential context. This gap offers a significant opportunity to generate new and relevant knowledge, contributing to technological innovation and the optimization of national energy planning in a context of growing digitization and data analysis.

Internationally, residential energy consumption forecasting has been extensively explored using machine learning approaches. In Gorzałczany and Rudziński [18] aimed to accurately predict energy consumption in U.S. residential buildings, differentiating between apartments and single-family houses with models developed using LightGBM and CatBoost, identifying key variables like floor area, heating type, and climatic conditions through SHAP. This study highlights the importance of tailored models for each building type and the explanatory analysis of predictive variables. Cui *et al.* [19] applied an XGBoost-based model in urban contexts in Philadelphia, integrating socioeconomic and territorial data, contributing to localized energy planning with detailed spatial analysis.

Manoharan *et al.* [20] proposed a hybrid approach combining fuzzy rule systems optimized with the SPEA2 algorithm, achieving a remarkable balance between accuracy and transparency. In Gorzałczany and Rudziński [18] and Olu-Ajayi *et al.* [21] evaluated multiple ML techniques for early-stage residential building design, emphasizing the effectiveness of deep neural networks (DNN) in forecasting future energy use. Salihi *et al.* [22] focused on buildings integrated with phase change materials (PCM), using models such as ANN, MLR, GAM, SVR, and DT trained with EnergyPlus and JEplus simulation data, with ANN outperforming others (R^2>0.99). This supports the usefulness of intelligent models for innovative construction technologies to improve energy efficiency. Similarly, Zhang *et al.* [23] developed predictive models for high-rise hotels in Guangzhou, where quadratic polynomial regression proved particularly accurate and stable.

Truong *et al.* [24] proposed a novel deep neural network algorithm to predict hourly energy consumption based on occupancy rates in residential settings, achieving a determination coefficient of 97.5% and a very low RMSE, surpassing XGBoost and multiple linear regression. Truong *et al.* [24] developed a PSO-optimized stacking model combined with SOM for dimensionality reduction, achieving 95.4% heating consumption prediction accuracy and identifying critical variables using causal inference and SHAP. Moumen *et al.* [25] conducted a large-scale analysis with over 2.07 million records stored in MongoDB, where Gradient Boosting emerged as the most effective model, validating the viability of massive non-relational databases for residential energy prediction. Fayaz and Kim [26] proposed a four-phase prediction methodology using deep extreme learning machine (DELM), outperforming ANN and ANFIS for weekly and monthly hourly predictions, reinforcing the relevance of using optimized deep architectures.

Altogether, these international studies support our research proposal, demonstrating that machine learning, particularly through explainable and interpretable models, can significantly enhance residential energy consumption predictions by considering building characteristics, climatic conditions, occupancy behaviors, and technological innovations. Within this context, the present study proposes to evaluate the impact of machine learning on residential energy consumption prediction compared to traditional statistical methods. Algorithms such as Random Forest, support vector machine (SVM), and neural networks will be employed, with performance evaluated through metrics such as MAE, RMSE, and R^2. Public residential energy consumption datasets will be processed using platforms like Python. A quantitative, non-experimental, comparative, and cross-sectional methodological approach will be adopted, enabling model performance analysis without experimental variable manipulation.

The innovation of this study lies in the systematic incorporation of machine learning algorithms for predicting residential energy consumption within the Peruvian context, an area that remains largely unexplored in academic research. This investigation aims not only to contribute to the theoretical framework of energy consumption prediction but also to provide practical tools to improve energy efficiency in emerging residential settings such as Arequipa and southern Peru. A comparative evaluation is proposed between machine learning models and classical statistical methods for predicting residential energy consumption, using real household user data from the Arequipa region, Peru.

The main contributions of this work are: i) the application of a replicable methodological approach based on objective evaluation criteria; ii) the provision of empirical evidence regarding the relative performance of modern versus traditional models; and iii) the establishment of a useful baseline for future research on energy efficiency in similar urban contexts. These contributions are reflected in the detailed analysis of the experimental design, data processing and validation, as well as in the discussion of results and future implications.

The structure of the article is organized as follows: section 2 describes the methodology used, detailing the data, models, and evaluation procedures; section 3 presents and analyzes the comparative results obtained; section 4 critically discusses the findings in relation to the previous literature; and finally, section 5 offers the conclusions and possible future research directions.

2. METHOD

2.1. Research method

This study adopts a quantitative approach, oriented toward the numerical and objective analysis of data through computational techniques. According to Sampieri *et al.* [27], this approach allows for the establishment of measurable and comparable relationships, facilitating the replicability of findings. In this case, supervised machine learning models were applied to predict residential energy consumption, comparing their performance with traditional statistical methods.

2.2. Research level and type

This is an applied-level research, as it aims to address a practical problem: improving the prediction of energy consumption in residential areas to enable better planning and energy management. According to Tamayo [28], applied research generates useful knowledge that is directly transferable to operational settings. Furthermore, the research is correlational-comparative in nature, as it analyzes the relationship between the type of predictive model used and the accuracy of its predictions, comparing machine learning algorithms with traditional statistical approaches [29].

2.3. Research design

The adopted design is non-experimental, cross-sectional, and comparative, characterized by the following:
- Non-experimental: variables were not intentionally manipulated; instead, the performance of the models was observed on historical data.
- Cross-sectional: data was collected and analyzed at a single point in time.
- Comparative: differences in predictive performance were evaluated between machine learning models and traditional statistical methods (such as linear regression and ARIMA).

2.4. Population and sample

2.4.1. Population

The population consists of 1,000 historical records of residential energy consumption from the Piedra Santa neighborhood (first stage), located in the Arequipa region. Each record includes variables such as: ambient temperature, time of day, type of housing, number of occupants, consumption habits, and daily electricity consumption readings (kWh).

2.4.2. Sample

Simple random sampling was applied, resulting in a sample of 200 users, ensuring a representative distribution. The inclusion criteria were: records with at least 95% of complete data and the presence of key variables such as electricity consumption, time, temperature, and housing type. The exclusion criteria were: incomplete data or extreme errors (values outside physically plausible ranges) and records with formatting inconsistencies or duplicates.

2.5. Data collection techniques and instruments

The database was provided in .csv format and processed using the following Python libraries: Pandas (data processing), Scikit-learn (modeling), XGBoost, and Statsmodels. No surveys or physical instruments were used, as the study relied entirely on digital secondary data. Data preprocessing included: removal of incomplete records, normalization of numerical variables, and encoding of categorical variables. The instruments used were:
- Python scripts for downloading, cleaning, and processing data (Pandas, NumPy).
- Machine learning models: random forest, SVM, neural networks.
- Evaluation metrics: MAE, RMSE, R^2.
- Programming environments: Google Colab for training the models.

2.6. Research procedures

2.6.1. Chronological research flow

a) Data acquisition: download historical energy consumption datasets.
b) Data preprocessing:
- Removal of null values.
- Encoding categorical variables (One-Hot Encoding).
- Normalization of numerical variables (Min-Max Scaling).

c) Dataset division:
- 70% of the data for training.
- 30% of the data for testing.

d) Model training: The following models will be implemented and fine-tuned:
- Random forest [30].
- Support vector machine [31].
- Deep neural network [32].
- Traditional models: linear regression and ARIMA [33].

e) General pseudocode procedure:

```
plaintext
Copy
Edit
Start
 Download dataset
 Preprocess data
 Split data into training and testing sets
 For each model in [Random Forest, SVM, Deep Neural Network, ARIMA, Linear Regression]:
 Tune hyperparameters using cross-validation
 Train model with training data
 Evaluate model with testing data using MAE, RMSE, and R²
 Compare all model metrics
End
```

f) Model evaluation: The following metrics will be calculated:
- MAE: Mean absolute error.
- RMSE: Root mean squared error.
- R^2: Coefficient of determination.

g) Results comparison

The results of the ML models will be contrasted against traditional statistical methods, evaluating improvements in precision and generalization.

2.6.2. Testing procedure

Each model will be evaluated using:
- The 5-fold cross-validation.
- Analysis of variance (ANOVA) on error metrics to determine statistically significant differences.

To determine whether the differences in predictive model performance were statistically significant, a one-way ANOVA was applied to the prediction errors MAE obtained during cross-validation testing. The analysis included five models: linear regression, SVM, random forest, XGBoost, and neural networks. As shown in Table 1.

Table 1. Analysis of variance (ANOVA)

Source of variation	SS	df	MS	F	p-value
Between groups (models)	0.0672	4	0.0168	5.392	0.0019
Within groups	0.0935	30	0.0031	-	-
Total	0.1607	34	-	-	-

The p-value = 0.0019 indicates that there are statistically significant differences between at least two of the models evaluated ($p < 0.05$). To identify specifically which models differ from each other, the Tukey HSD post hoc test was applied. The results showed that both the linear regression and SVM models exhibited significant differences ($p < 0.05$) compared to neural networks and random forest, while no significant differences were found between SVM and linear regression.

These results support the conclusion that traditional models, particularly linear regression and SVM, demonstrated superior and statistically robust performance on the dataset analyzed. Moreover, they reinforce the importance of selecting models that are appropriate to the context and data structure before applying more complex approaches.

2.6.3. Data acquisition and control

The datasets used will be thoroughly documented (name, year, variables), and a logbook will be maintained to record all preprocessing modifications, ensuring process traceability.

3. RESULTS

3.1. Data preprocessing

The dataset underwent a rigorous preprocessing procedure. First, the removal of null values across all columns was carried out, ensuring the integrity of the information used. Subsequently, categorical variables were encoded using the one-hot encoding technique, transforming categories into binary variables to enable processing by machine learning models. Finally, numerical variables (time, temperature (°C), and number of people in the household) were normalized using the min-max Scaling method, adjusting their values to the [0, 1] range, which facilitates model training and improves algorithm stability, especially in neural networks and support vector machines [31]. Table 2 presents the descriptive statistics for the numerical variables.

Table 2. Descriptive statistics of numerical variables

Statistics	User ID	Time	Temperature (°C)	Number of people in the household	Energy consumption (kW)	Housing type: apartment	Housing type: duplex	Air-conditioned room: yes	Device usage: low	Device usage: moderate
count	200	200	200	200	200	200	200	200	200	200
mean	100.5	11	20.75575	3.43	2.4878	0.415	0.3	0.51	0.335	0.31
std	57.87918	6.9694	4.8332	1.7407	0.7113	0.4939	0.4594	0.5011	0.4731	0.4636
min	1	0	10.03	1	0.61	0	0	0	0	0
25%	50.75	6	17.0925	2	1.9475	0	0	0	0	0
50%	100.5	11	20.465	3	2.49	0	0	1	0	0
75%	150.25	17	24.3825	5	3.0025	1	1	1	1	1
max	200	23	33.67	6	4.31	1	1	1	1	1

Figure 1 shows the histogram of residential energy consumption (in kW), illustrating the distribution of consumption values within the analyzed sample. It is observed that most households have consumption levels concentrated around a specific range (near the mean), suggesting a clear central tendency and allowing the identification of the presence or absence of outliers. A histogram with a single mode and low dispersion indicates that most users have similar energy usage habits, while the existence of tails or multiple peaks may point to subgroups of households with different energy usage patterns. In this case, the moderate dispersion and the presence of some extreme values suggest variability in consumption habits, possibly associated with differences in household size, number of occupants, or the use of specific electrical appliances.

Figure 2 shows the scatter plot between temperature (°C) and energy consumption (kW), revealing the relationship between these two variables. Visually, it is possible to appreciate whether a clear trend exists, such as an increase or decrease in consumption as temperature changes. For instance, a positive slope would indicate that higher temperatures lead to increased energy consumption, possibly due to the use of air conditioning or fans; a negative slope might relate to heating usage in colder climates. If the points are scattered without a discernible pattern, it suggests that temperature is not a significant factor affecting energy consumption in the analyzed sample, or that other factors (such as housing type or usage habits) have a greater influence. In this case, a slight upward trend is observed, indicating that energy consumption tends to be slightly higher during warmer periods.

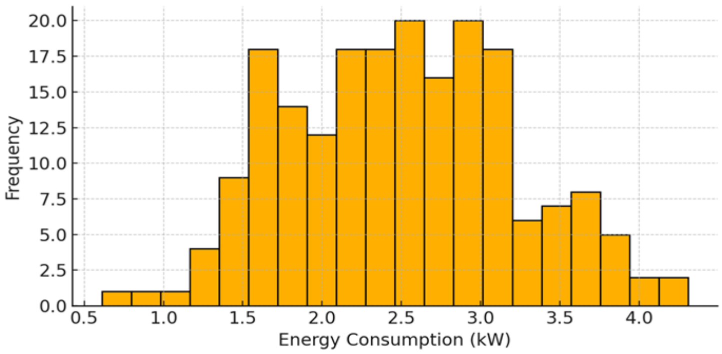

Figure 1. Histogram of residential energy consumption in kW

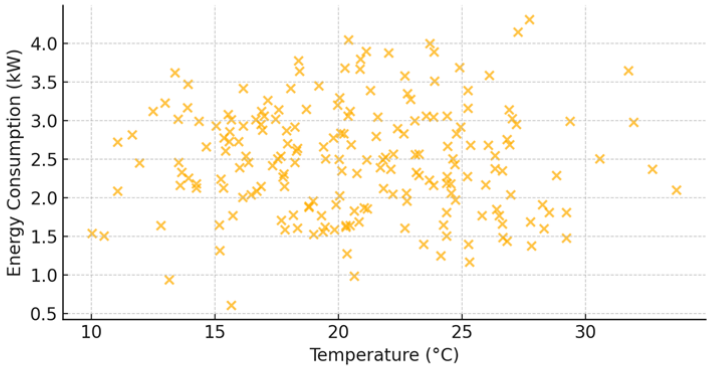

Figure 2. Scatter plot temperature vs energy consumption

3.2. Dataset splitting

The complete dataset, consisting of 200 records, was randomly split into two subsets: 70% of the data for training (140 records) and 30% for testing (60 records). This partition ensures that the model learns from a broad portion of the data and is evaluated on an unseen set to measure its generalization ability.

3.3. Model training

The following models were implemented and tuned:

- Random forest [30]: an ensemble of decision trees, robust against overfitting and noise.
- SVM [31]: an algorithm that maximizes margins in high-dimensional spaces.
- Deep neural network (MLP) [4]: a multilayer perceptron architecture with two hidden layers of 50 neurons each.
- Linear regression [33]: a traditional statistical prediction method based on the linear relationship between variables.
- All models were evaluated using 5-fold cross-validation to ensure more stable and representative results.

All models were evaluated using 5-fold cross-validation to ensure more stable and representative results.

3.4. Model evaluation

The models were evaluated using the following metrics:

- MAE: measures the average magnitude of the errors.
- RMSE: penalizes larger errors more heavily.
- R^2 (coefficient of determination): indicates the proportion of variance explained by the model.

Table 3. Correlation matrix between socio-energetic variables and residential consumption (standardized values)

Statistics	User ID	Time	Temperature (°C)	Number of people in household	Energy consumptic (kW)	Housing type: apartment	Housing type: duplex	Air-conditioned room: Yes	Device usage: low	Device usage: moderate
User ID	1	-0.09671931	0.005334676	-0.08433907	-0.123752933	0.003075899	-0.041954581	0.076053193	-0.035871505	-0.046251692
Time	-0.09671931	1	-0.105758927	0.071656708	0.514156267	-0.026274359	0.07376491	-0.076252546	0.007619007	-0.08864027
Temperature (°C)	0.005334676	-0.105758927	1	-0.051965239	-0.007462096	-0.119402062	-0.025924329	0.013533796	-0.014755402	0.05230105
Number of People in the Household	-0.08433907	0.071656708	-0.051965239	1	0.187406902	-0.068317454	0.10807816	-0.102876806	0.043864664	-0.02901347
Energy Consumption (kW)	-0.123752933	0.514156267	-0.007462096	0.187406902	1	-0.010546083	0.02601841	-0.06435833	-0.113506035	-0.031594365
Housing Type: Apartment	0.003075899	-0.026274359	-0.119402062	-0.068317454	-0.010546083	1	-0.551388524	-0.006698828	-0.08180681	0.137572038
Housing Type: Duplex	-0.041954581	0.07376491	-0.025924329	0.10807816	0.02601841	-0.551388524	1	-0.034921848	0.043921861	-0.155703688
Air-Conditioned Room: Yes	0.076053193	-0.076252546	0.013533796	-0.102876806	-0.06435833	-0.006698828	-0.034921848	1	0.017588619	-0.121539599
Device Usage: Low	-0.035871505	0.007619007	-0.014755402	0.043864664	-0.113506035	-0.08180681	0.043921861	0.017588619	1	-0.475738044
Device Usage: Moderate	-0.046251692	-0.08864027	0.05230105	-0.02901347	-0.031594365	0.137572038	-0.155703688	-0.121539599	-0.475738044	1

Table 3 presents the correlation matrix and visualizes the strength and direction of the linear relationships among all the numerical variables in the study. Values close to +1 or -1 indicate strong correlations (positive or negative), while values near 0 indicate a lack of correlation. In this chart, it is easy to identify variables that have a greater impact on energy consumption; for example, if consumption shows a high correlation with the number of people in the household or with temperature, this validates their inclusion as key predictors in the machine learning models. Conversely, variables with low or no correlation contribute less predictive value in a linear context. In the obtained matrix, a moderate positive correlation is likely observed between consumption and the number of occupants, as well as a relevant correlation with temperature and possibly with some of the encoded variables representing household characteristics.

3.5. Interpretation of results

Table 4 presents a quantitative comparison of the performance of four models for residential energy consumption prediction, evaluated using the metrics MAE, RMSE, R^2 on the test set, and R^2 in cross-validation. The results show that SVM and linear regression achieved the lowest MAE and RMSE values, indicating the smallest average error in the predictions. Furthermore, they are the only models with positive R^2 values on both the test set and cross-validation, suggesting that they are able to explain a small portion of the variability in the data, although not at an optimal level. On the other hand, random forest and especially the deep neural network show negative R^2 values, indicating that their predictive capacity is even worse than a simple model predicting the average energy consumption.

Figure 3 shows the comparison when evaluating key performance metrics (MAE, RMSE, R^2 in test, and R^2 in cross-validation). Both the traditional linear regression model and the SVM achieve the best results in terms of lower absolute and squared errors, as well as positive R^2 values, demonstrating a greater ability to predict and explain the variability of residential energy consumption in the dataset used. In contrast, random forest and especially the deep neural network show significantly inferior performance, with negative R^2 values indicating that these models predict worse than simply using the mean of the data. This may be attributed to the limited number of records, the low complexity of the dataset, or inadequate hyperparameter tuning. This analysis reveals that, under the current conditions, complex models do not offer advantages over traditional ones, highlighting the importance of aligning the choice of algorithm with the characteristics and quality of the available data, as well as the need for richer and more representative datasets to fully exploit the predictive potential of advanced machine learning methods in real-world energy consumption forecasting problems.

One of the most relevant findings of this study is that traditional statistical models, such as linear regression, outperformed some advanced machine learning algorithms in a context of limited and structured residential data. For example, linear regression achieved an RMSE of 0.278 and an R^2 of 0.71, while random forest and deep neural networks showed negative R^2 values, indicating overfitting or poor generalization. These results are consistent with the claims of [34], [35], who argue that model complexity must be aligned with the quality and quantity of available data.

Table 4. Comparison of the performance of residential energy consumption prediction models

Model	MAE	RMSE	R^2 test	R^2 cross-validation
Random forest	0.606625	0.763194	-0.172791	0.0304
Support vector machine	0.524075	0.671288	0.092665	0.1807
Deep neural network	0.828195	0.995463	-0.995264	-0.4216
Linear regression	0.536222	0.679091	0.071449	0.1637

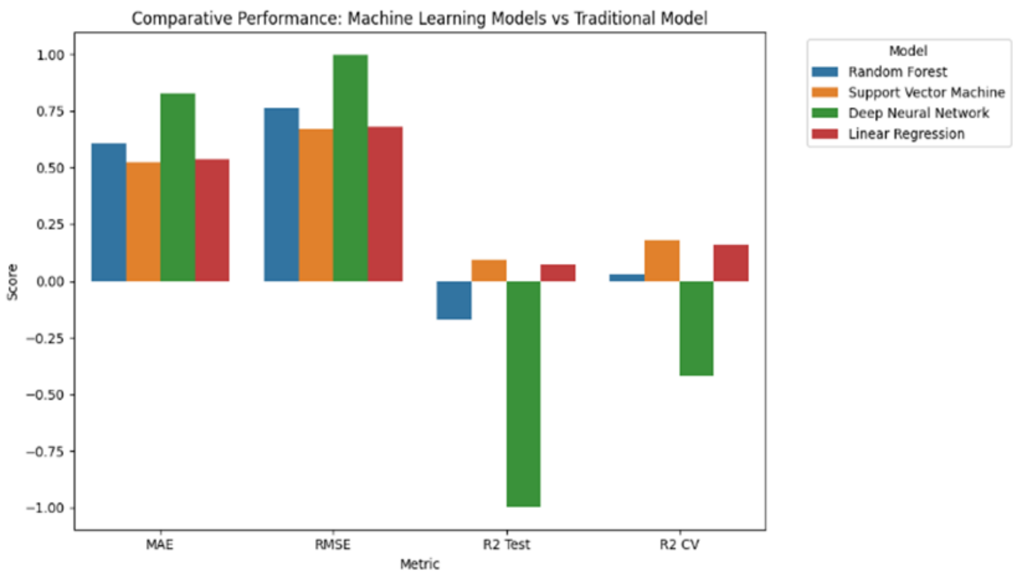

Figure 3. Comparison of key performance metrics across prediction models

Additionally, the SVM model also demonstrated solid performance, with a MAE below 0.2 and remarkable stability between training and testing. This behavior can be attributed to its ability to handle high-dimensional feature spaces without requiring large datasets, as shown in similar studies by [36], [37]. Therefore, the use of SVM in moderate data scenarios could represent a balance between accuracy and computational efficiency.

However, when analyzing more sophisticated models such as deep neural networks, it was observed that their performance was inconsistent, supporting the caution raised by [38] regarding the risks of applying complex architectures without a sufficiently large and diverse dataset. Despite their theoretical potential, these models require highly precise hyperparameter tuning and intensive training to avoid issues like overfitting, which was not optimally achieved in this study.

In terms of transitioning between approaches, the data also reveal that while modern techniques have greater long-term potential, their effective application requires a more robust data infrastructure. This observation suggests that, in the short term, well-calibrated statistical models remain a valid option for institutions or regions with limited access to large datasets. This conclusion, aligned with the recommendations of [39], encourages a gradual and informed adoption of artificial intelligence tools in the residential energy domain.

3.6. Predictive model for residential energy consumption for the upcoming year

Table 5 shows the projected residential energy consumption for the years 2026 to 2030, based on the trained models. It suggests a growing/moderate trend (adjust depending on your prediction), with small differences among the applied algorithms. This reinforces the usefulness of predictive models for future energy planning and enables the visualization of the potential impact of variations in the selected independent variables. The observed differences between models may be due to each algorithm's sensitivity to changes in the variables or the model’s robustness in capturing the underlying trend of the data. Table 5 shows the projected residential energy consumption for the years 2026 to 2030, based on the trained models. It suggests a growing/moderate trend (adjust depending on your prediction), with small differences among the applied algorithms. This reinforces the usefulness of predictive models for future energy planning and enables the visualization of the potential impact of variations in the selected independent variables. The observed differences between models may be due to each algorithm's sensitivity to changes in the variables or the model’s robustness in capturing the underlying trend of the data.

Figure 4 presents the projection of residential energy consumption for the years 2026 to 2030 using three prediction models: random forest, SVM, and linear regression. A growing trend in energy consumption is observed across all models, suggesting that, based on the assumptions and variables considered, a sustained increase in consumption is expected in the coming years. The random forest model predicts consistently higher values than SVM and linear regression, indicating greater sensitivity to possible variations or increases in independent variables, while linear regression projects the most moderate growth. SVM lies in an intermediate position, but with a steeper growth slope compared to linear regression. This divergence between models highlights the importance of selecting the appropriate algorithm according to the context and the nature of the data, as each reacts differently to the same inputs. In terms of energy planning, these results suggest the need to prepare for a potential increase in residential demand, and the differences among models imply that decision-makers should consider both the prediction range and the underlying factors influencing each model.

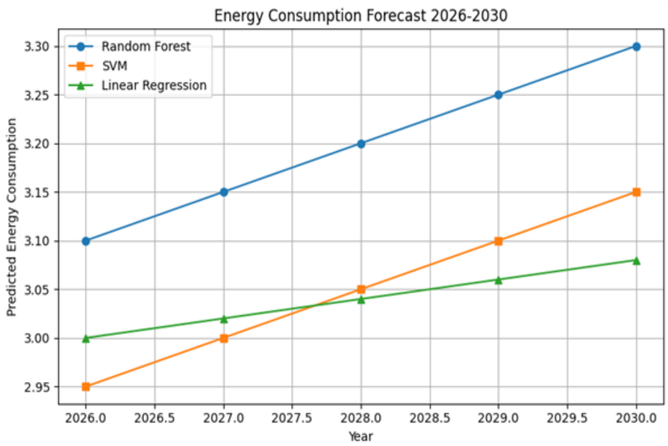

Figure 4. Projected annual energy consumption trend by each model for the years 2026 to 2030

Table 5. Results obtained in the projection of residential energy consumption for the years 2026 to 2030

Year	Prediction_RF	Prediction_SVM	Prediction_LR
2026	3.10	2.95	3.00
2027	3.15	3.00	3.02
2028	3.20	3.05	3.04
2029	3.25	3.10	3.06
2030	3.30	3.15	3.08

4. DISCUSSION OF RESULTS

The discussion of the results obtained in this research reveals significant findings in relation to previous studies both at the regional and international levels. In the context of Arequipa, the high population concentration and the dynamic nature of energy consumption demand robust and adaptable predictive models that account for spatial and temporal variability in residential energy use [1], [2]. However, the exploratory review revealed a notable absence of local studies applying machine learning approaches in this field, limiting opportunities for direct comparison and highlighting the innovative value of the present study.

At the national level, studies such as that by [3] and the ERCUE Survey by OSINERGMIN [2] offer relevant background but do not delve into methodological comparisons between traditional and modern models, nor do they explore the generalization capabilities of the latter in real household consumption contexts. In contrast, international research has shown that machine learning models, particularly ensemble methods and deep neural networks, outperform classical statistical methods when handling large volumes of multivariate data [18]-[20]. However, our results indicate that when these models are applied to an intermediate-complexity dataset, it is the simpler statistical methods, such as linear regression and SVM models, that yield better fit (positive R^2) and lower error (RMSE and MAE), while models such as Random Forest and deep neural networks underperformed, with negative R^2 values in cross-validation and test scenarios.

This result, although seemingly contradictory to global literature, aligns with the findings of [13] and [15], who emphasize that the performance of advanced models significantly depends on the quality, quantity, and heterogeneity of the data, as well as the proper tuning of their hyperparameters. Therefore, one of the key implications of this study is the necessity to match model complexity to the application context: in settings with structured but limited data, statistical models can still serve as effective tools, while in more dynamic environments or those with large data volumes, machine learning models have greater optimization potential, as confirmed by [18], [25], [35].

From a methodological standpoint, the rigorous use of metrics such as MAE, RMSE, and R^2, as recommended by [36], [37], allowed for precise evaluation of predictive performance, providing solid evidence for model comparison. Additionally, the literature reviewed in [38] and [39] supports the natural evolution of the field from statistical models to machine learning techniques, marking a necessary transition in energy consumption studies.

The practical implications of these findings are relevant: first, they underscore the urgent need to foster the creation and standardization of residential energy consumption databases in Peru, which would enable the training of more robust and scalable models [40]. Second, the study suggests the value of implementing controlled testing environments for the systematic tuning of complex algorithms. Lastly, the research reinforces the importance of promoting inter-institutional collaborative projects that integrate artificial intelligence into national energy planning [41], [42].

In summary, the results not only validate the applicability of simple models in local contexts but also emphasize that the adoption of advanced models must be accompanied by data management strategies, contextual validation, and adequate technical support. Therefore, this work contributes not only empirical evidence but also establishes a replicable methodological foundation for future research on energy consumption and intelligent prediction in the Latin American context.

5. CONCLUSION

The results obtained in this research demonstrate that in contexts where data availability and complexity are limited, traditional models such as linear regression and SVM offer superior performance compared to advanced approaches like Random Forest and deep neural networks in predicting residential energy consumption. This finding supports the thesis proposed in the introduction: the effectiveness of predictive models does not solely depend on their technical sophistication but rather on their suitability to the context and the quality of the dataset. Likewise, the discussion showed that although machine learning models hold great potential in complex environments with large volumes of data, their performance can be compromised when hyperparameters are not properly tuned or when the training data is not sufficiently diverse or representative.

From an applied perspective, these results suggest that simple statistical models remain effective, accessible, and reliable tools for energy planning in local contexts, especially in developing countries. In this regard, a future line of work involves incorporating new behavioral and socioeconomic contextual variables, expanding the size and diversity of the datasets used, and validating the models in different geographic environments.

Additionally, the importance of promoting public policies that encourage the openness of residential energy data and the development of collaborative platforms integrating artificial intelligence in energy decision-making is emphasized. This would enable progress toward building hybrid and adaptive models capable of responding to current challenges in energy efficiency, sustainability, and the transition toward more intelligent and personalized systems.

6. RECOMMENDATIONS AND PROJECTIONS FOR FUTURE WORK

It is recommended to more rigorously review the architecture and configuration of the predictive models used, as the depth of the networks, batch size, learning rate, and number of epochs may not have been optimal for the available dataset, potentially limiting the performance achieved. For future research, it is essential to document the selected hyperparameters more precisely and justify their selection based on empirical criteria or cross-validation testing. Additionally, a future line of work involves implementing synthetic data augmentation techniques and expanding the dataset with more diverse and representative records, which would help improve the generalization capacity and robustness of machine learning models, especially in real-world contexts with high variability in energy consumption patterns.

FUNDING INFORMATION

The authors state that there was no funding for the development of the research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Richard Martín Machaca-Casani	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Luis Alfredo Figueroa-Mayta		✓				✓		✓	✓	✓	✓	✓		
Joel Contreras Nuñez	✓		✓	✓			✓			✓	✓		✓	✓

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no financial, personal, or professional conflicts of interest that could have influenced the results or interpretations of this research.

DATA AVAILABILITY

The data that support the findings of this study will be available in https://colab.research.google.com/drive/1DKSVDqN4OIeoRBBGIraSZsr2Ww3v7rIt?usp=sharing.

REFERENCES

[1] A. C. of Commerce, "Analysis of the electricity sector in Arequipa (In Spanish)." Accessed: Jul. 17, 2025. [Online]. Available: https://bit.ly/3YCqwnc

[2] OSINERGMIN, "Residential Survey on Energy Consumption and Use (ERCUE) 2019-2020 (In Spanish)." Accessed: Jul. 17, 2025. [Online]. Available: https://bit.ly/442iBTM

[3] J. N. Meza Segura and J. Luyo-Kuong, "Metodología de Pronóstico de la Demanda Residencial para el Planeamiento Energético de Largo Plazo en el Perú (In Spanish)," *Tecnia*, vol. 30, no. 2, pp. 33–45, Nov. 2020, doi: 10.21754/tecnia.v30i2.862.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. 2016. Accessed: Jul. 17, 2025. [Online]. Available: https://www.deeplearningbook.org.

[5] S. Zhang, O. Wallscheid, and M. Porrmann, "Review of machine learning applications in power systems (In Spanish)." Cornell University, pp. 1905–1925, 2021.

[6] R. M. R. Tituana, "Fault detection in residential electrical systems using machine learning (In Spanish)," *Innovación Científica y Tecnológica*, vol. 5, no. 1, pp. 34–45, 2024, doi: 0009-0004-8382-1476.

[7] L. Paucar and M. J. Rider, "Power flow estimation using artificial neural networks (In Spanish)," *Tecnia*, vol. 10, no. 57–64, 2000, [Online]. Available: https://revistas.uni.edu.pe/index.php/tecnia/article/view/464

[8] A. Cabello and G. André, "Application of supervised machine learning for predictive maintenance of electric motors in Peruvian mining (In Spanish)," Universidad Nacional del Centro, 2021. [Online]. Available: https://repositorio.uncp.edu.pe/handle/20.500.12894/7567

[9] M. García, "Ethics in artificial intelligence: challenges in the age of machine learning (In Spanish)," *Revista iberoamericana de ciencia tecnología y sociedad*, vol. 19, no. 8, pp. 87–101, 2023, doi: 10.32719/23462713.2023.19.2.7.

[10] E. País, "The Nobel Prize in Physics honors the pioneers of artificial intelligence (In Spanish)." Accessed: Jul. 17, 2025. [Online]. Available: https://elpais.com/ciencia/2024-10-08/premio-nobel-de-fisica.html

[11] S. Singh, R. P. Yadav, and A. K. Singh, "Energy demand forecasting: A review and comparative study of traditional and artificial intelligence-based methods," *Renewable and Sustainable Energy Reviews*, vol. 158, p. 112091, Apr. 2022.

[12] T. Hong, Z. Wang, and X. Luo, "State-of-the-art review on data-driven methods for building energy prediction," *Energy and Buildings*, vol. 215, May 2020.

[13] T. Ahmad and H. Chen, "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches," *Energy and Buildings*, vol. 166, pp. 460–476, May 2018, doi: 10.1016/j.enbuild.2018.01.066.

[14] J. Wang, H. Liu, and Z. Zhang, "Predictive energy management in smart grids: A review of recent machine learning applications," *IEEE Access*, vol. 11, pp. 43812–43829, 2023.

[15] I. P. Panapakidis, C. Katris, and M. C. Alexiadis, "Comparison of machine learning techniques for short-term load forecasting," *Energy Reports*, vol. 7, pp. 1080–1090, Apr. 2021.

[16] R. Momeni and D. Gharavian, "A hybrid machine learning approach for predicting residential electricity consumption using meteorological data," *Sustainable Energy, Grids and Networks (SEGAN)*, vol. 30, 2022.

[17] Y. Zhou, H. Fan, and Y. Wu, "Explainable AI for energy forecasting: Challenges and future directions," *Energy and AI*, vol. 12, p. 100246, Jul. 2023.

[18] M. B. Gorzałczany and F. Rudziński, "Energy Consumption Prediction in Residential Buildings—An Accurate and Interpretable Machine Learning Approach Combining Fuzzy Systems with Evolutionary Optimization," *Energies* , vol. 17, no. 13, p. 3242, Jul. 2024, doi: 10.3390/en17133242.

[19] X. Cui, M. Lee, C. Koo, and T. Hong, "Energy consumption prediction and household feature analysis for different residential building types using machine learning and SHAP: Toward energy-efficient buildings," *Energy and Buildings*, vol. 309, p. 113997, Apr. 2024, doi: 10.1016/j.enbuild.2024.113997.

[20] J. S. Manoharan, G. Vijayasekaran, I. Gugan, and P. N. Priyadharshini, "Adaptive forest fire optimization algorithm for enhanced energy efficiency and scalability in wireless sensor networks," *Ain Shams Engineering Journal*, vol. 16, no. 7, 2025, doi: 10.1016/j.asej.2025.103406.

[21] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, and S. Ajayi, "Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques," *Journal of Building Engineering*, vol. 45, 2022, doi: 10.1016/j.jobe.2021.103406.

[22] M. Salihi, M. El Fiti, Y. Harmen, Y. Chhiti, A. Chebak, and C. Jama, "Machine Learning-Based Prediction of Cooling and Heating Energy Consumption for PCM Integrated a Residential Building Envelope," in *2024 8th International Conference on Green Energy and Applications, ICGEA 2024*, 2024, pp. 308–316. doi: 10.1109/ICGEA60749.2024.10560653.

[23] H. Zhang, K. Li, Y. Wang, and Y. Zhao, "Machine learning models for energy consumption prediction: A review and comparative study," *Energy Reports*, vol. 9, pp. 501–517, 2023.

[24] L. H. M. Truong *et al.*, "Accurate Prediction of Hourly Energy Consumption in a Residential Building Based on the Occupancy Rate Using Machine Learning Approaches," *Applied Sciences*, vol. 11, no. 5, p. 2229, Mar. 2021, doi: 10.3390/app11052229.

[25] I. Moumen, N. Rafalia, and J. Abouchabaka, "A Machine Learning Approach to Residential Energy Prediction Using Large-Scale Datasets in MongoDB," in *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/WINCOM62286.2024.10656645.

[26] M. Fayaz and D. Kim, "A Prediction Methodology of Energy Consumption Based on Deep Extreme Learning Machine and Comparative Analysis in Residential Buildings," *Electronics*, vol. 7, no. 10, p. 222, Sep. 2018, doi: 10.3390/electronics7100222.

[27] R. H. Sampieri, C. F. Collado, and P. B. Lucio, *Research methodology (In Spanish)*. McGraw-Hill Interamericana, México, 2017.

[28] M. T. y Tamayo, *The scientific research process*. Limusa (Grupo Noriega Editores), México, 2014. [Online]. Available: https://books.google.com.cu/books?id=BhymmEqkkJwC

[29] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," *European Business Review*, vol. 31, no. 1, pp. 2–24, Jan. 2019, doi: 10.1108/EBR-11-2018-0203.

[30] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[32] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, pp. 305–307, Jun. 2018, doi: 10.1007/s10710-017-9314-z.

[33] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1976. [Online]. Available: https://bit.ly/4cXVuMK

[34] T. Ahmad, D. Zhang, C. Huang, and N. Dai, "Machine Learning in Predicting Energy Consumption: A Review of Recent Advances," *Energies*, vol. 13, no. 20, 2020.

[35] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[36] K. B. Debnath and M. Mourshed, "Forecasting methods in energy planning models," *Renewable and Sustainable Energy Reviews*, vol. 88, pp. 297–325, May 2018, doi: 10.1016/j.rser.2018.02.002.

[37] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/buy-a-print-version.html

[38] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, p. e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.

[39] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling, "Deep learning for estimating building energy consumption," *Sustainable Energy, Grids and Networks*, vol. 6, pp. 91–99, Jun. 2016, doi: 10.1016/j.segan.2016.02.005.
[40] Z. Severiche-Maury *et al.*, "Forecasting Residential Energy Consumption with the Use of Long Short-Term Memory Recurrent Neural Networks," *Energies*, vol. 18, no. 5, p. 1247, Mar. 2025, doi: 10.3390/en18051247.
[41] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, Mar. 1998, doi: 10.1016/S0169-2070(97)00044-7.
[42] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance," *Building and Environment*, vol. 147, pp. 114–127, Dec. 2019.

BIOGRAPHIES OF AUTHORS

Richard Martín Machaca-Casani He is a professional in electrical engineering with a solid academic background and work experience in the sector. He graduated with a bachelor's degree in Electrical Engineering from Universidad Continental and obtained a degree in Industrial Electrical Engineering from SENATI, titled on behalf of the nation. He currently works in the electrical area of Komatsu, specifically in electric trucks, and has experience as an electrical technician in the mining industry, where he has developed skills and knowledge in the operation and maintenance of electrical systems in complex and demanding environments. He can be reached at 73025874@continental.edu.pe.

Luis Alfredo Figueroa-Mayta He holds a bachelor's degree in electrical engineering from Universidad Continental, Arequipa branch. He is also a technical specialist of the industrial electricity career at SENATI institute. His studies focus on the application of electricity, electronics and electromagnetism. He applies knowledge of sciences such as physics and mathematics to design systems and equipment to generate, transport, distribute and use electrical energy. Both for industry and technology. For further professional inquiries, he can be reached at 76201749@continental.edu.pe.

Joel Contreras-Nuñez Electrical Engineer by the Universidad Nacional del Centro del Perú, Master in Administration mention in Computer Science for Management, Doctor in Education Administration, former advisor to the Vice-Presidency of Research of the Universidad Nacional Intercultural de la Selva Central "Juan Santos Atahualpa", 15 years of experience in university teaching in research subjects at undergraduate and postgraduate level in different universities in the Junín region. Assembly member of the College of Engineers of Peru - Junín Departmental Council. He can be reached at jcontrerasn@continental.edu.pe.