

Received 9 October 2025, accepted 8 November 2025, date of publication 11 November 2025,
date of current version 26 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3631672

RESEARCH ARTICLE

EdgeAI-Powered Hybrid ESN-GRU Model for High-Accuracy and Efficient Short-Term Load Forecasting in Smart Grids

SUHIB BANI MELHEM¹, MUHAMMED GOLEC², SAED ALRABAE³,
MUATH ALSHAikh¹, AND MURAT UYAR⁴

¹Department of Cybersecurity, College of Engineering, Al Ain University, Abu Dhabi, United Arab Emirates

²Department of Computer Engineering, Boğaziçi University, Istanbul, Türkiye

³Information Systems and Security, College of IT, United Arab Emirates University, Al Ain, United Arab Emirates

⁴Electrical and Electronics Engineering Department, Bursa Uludağ University, Nilüfer, Türkiye

Corresponding authors: Saed Alrabae (salrabae@uaeu.ac.ae) and Muhammed Golec (muhammed.golec@bogazici.edu.tr)

ABSTRACT With the widespread use of renewable energy sources (RES) in the smart grid, the next generation power system, short-term load forecasting (STLF) is of critical importance in grid stability and energy optimization. Traditional STLF models include issues such as high computational cost, dependency on cloud infrastructure, and latency issues, which are undesirable for real-time energy management. To solve these issues, the EdgeAI paradigm, which combines edge computing and artificial intelligence (AI), can be a promising solution. EdgeAI reduces the dependency on cloud-based systems by processing data close to the data source, offering advantages such as low latency and low bandwidth. Thus, it increases the response speed by processing data in real time, making it suitable for STLF applications. In order to benefit from all these advantages, the EdgeAI-driven Hybrid Echo State Network and Gated Recurrent Unit (ESN-GRU) model is introduced for real-time and efficient STLF in this paper. ESN-GRU combines the fast training capabilities of ESN and sequential learning capabilities of GRU, and offers the advantages of high prediction performance and low latency inference in STLF. Experimental results show that the proposed model improves the R^2 score by 2.5% and significantly reduces the MAPE value from 129 to 0.101 compared to existing models. Benchmark results in edge environment prove that ESN-GRU provides up to 79% and 92% faster inference compared to state-of-the-art methods.

INDEX TERMS Gated recurrent unit (GRU), energy efficiency, smart grid, artificial intelligence, edge computing.

I. INTRODUCTION

Energy management and optimization have become a very critical issue with the diversification of Renewable Energy Sources (RES) and their integration into smart grids [1]. Solar and wind energy sources are more unpredictable due to their intermittent nature (change in wind speed, angle of sun rays, etc.) compared to traditional energy systems and it is more difficult to provide supply-demand balance [2]. In order to provide this balance, Short-Term Load Forecasting (STLF)

can increase grid resilience by making energy distribution and operating costs more efficient in cases where high forecast accuracy is provided.

In smart grids with complex production-consumption models, traditional statistical-based models are insufficient in load forecasting due to linear and stationary data assumptions [3]. For this reason, academic studies prefer Deep Learning (DL) and Machine Learning (ML) based approaches that capture patterns in complex datasets with superior performance [4]. In addition, there is an increasing tendency to integrate these DL & ML models used for STLF with cloud solutions that offer advantages such as

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang¹.

high processing power and dynamic scalability [5]. However, cloud-based systems bring additional challenges such as high latency, low security, inability to access resources in places with internet outages and bandwidth restrictions [6]. To overcome all these problems, the EdgeAI paradigm, which offers advantages such as low latency and the ability to work even in places without internet by bringing the processing power closer to the data source, can be promising. The EdgeAI paradigm is the combination of Edge Computing and Artificial Intelligence (AI) and offers lower processing power compared to cloud-based systems [7]. For this reason, the AI model to be used in STLF prediction should have low processing power and high prediction performance. In this paper, an EdgeAI-driven hybrid ESN-GRU model designed for resource-constrained edge environments is proposed. ESN-GRU leverages the computational capability of Echo State Networks (ESN) and the sequential modeling power of Gated Recurrent Units (GRU) to achieve minimum computational requirement and high accuracy.

A. CONTRIBUTION AND MOTIVATION

In smart grids, one of the frequently used methods in grid stability and energy optimization is STLF [8]. Existing approaches are inadequate in capturing nonlinear energy consumption patterns [9]. On the other hand, advanced DL-based methods used to solve this problem require high processing power and make it difficult to integrate with low-power edge devices [10], [11]. To solve this problem, the trend towards cloud-based systems with higher processing power has increased [12]. As highlighted in the introduction, cloud-based systems have latency and bandwidth limitations [13]. The EdgeAI paradigm addresses these limitations (enabling local processing) and is more suitable for real-time STLF applications [14]. However, due to the limited processing power due to the nature of edge computing, efficient and high-accuracy prediction models are needed [15].

This study presents a new approach to meet this need, the **Echo State Network + Gated Recurrent Unit (ESN-GRU)** hybrid model. ESN-GRU is a new energy-efficient and high-performance hybrid model for smart and sustainable energy management in smart grids. **The key contributions of this paper can be summarized as follows:**

- **General Purpose and Feature of the Model:** It is a new scalable and easy-to-integrate model designed for real-time deployment in smart grids.
- **Advantage of the Model for Edge Devices:** It is a model that provides fast inference and high prediction accuracy by reducing complexity for edge devices.
- **Experimental Validation:** Tests performed on real-world datasets confirm that the proposed model shows higher MAPE (a significant drop from 129 to 0.101) and R^2 score (2.5% increase) improvements compared to existing models.
- **Optimized for EdgeAI environments:** ESN-GRU is a lightweight model and can be deployed on edge devices

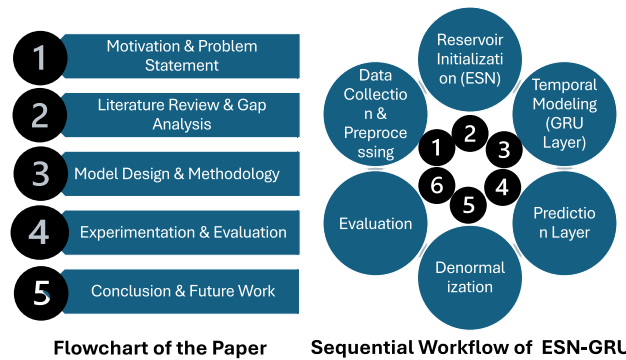


FIGURE 1. Flowchart representing the paper structure and the sequential workflow of the ESN-GRU model.

such as Raspberry Pi, reducing the dependency on cloud-based systems.

B. RESEARCH WORKFLOW AND PAPER STRUCTURE

The overall research workflow and paper structure are summarized in Figure 1. The flow of the paper is organized as follows;

Section I examines the background and motivation of EdgeAI-based studies for smart grid environments and the limitations of existing STLF methods. Section II examines existing academic studies by classifying them under the subheadings of Deep Learning, EdgeAI, and computationally optimized approaches. Section III introduces the proposed model, ESN-GRU, methodology, motivation, architectural design, and detailed mathematical formulation to the reader. Section IV provides the experimental setup, used datasets, evaluation metrics, and performance comparisons. Section V concludes the paper by highlighting research gaps for future studies.

Figure 1 shows the sequential workflow for the ESN-GRU model. Before the ESN-GRU model is trained, the data is first collected (step 1) and goes through preprocessing processes such as normalization. The aim of doing this is to increase the prediction performance of the model by cleaning the noisy and missing data in dataset. In step 2, the Reservoir Initialization process is performed so that the reservoir in the ESN model can sufficiently carry the effect of past data. In step 3, Temporal Modeling is performed to associate temporal models in GRU. In step 4, a prediction is made using the information that learns the time dependencies processed in the GRU layer. In step 5, the prediction results obtained using the normalized data (step 1) are denormalized and returned to the old scale. In step 6, the prediction performance of the ESN-GRU model is evaluated.

II. RELATED WORK

When the literature is examined, it is seen that the interest in STLF-based studies in RES studies has started to increase in order to provide a supply-demand balance and load optimization. In this section, we examine the solutions used in STLF applications by classifying them under the subheadings

TABLE 1. Comparison of ESN-GRU with existing works.

Study	Focus Area	Model Type	Data Size	Computational Cost	Edge Compatibility
[16]	BEF	LGBM	25.14 million data	M	L
[17]	OEF	ANN + RF + XGBoost	438 simulation data	M	M
[18]	STLF	CNN-LSTM + PSO	1 month building data (3600 m ²)	H	M
[19]	BEF	CNN + LSTM + MLP	43842 hourly data	M-H	L-M
[20]	ELF	GNN + LSTM + K-Medoids	6 months data (24 buildings)	M-H	L-M
[21]	STLF	KAN + RNN	10 different data sets	M-H	L-M
[22]	STLF	iTx + TL	4 data sets (UCI, PRECON, UC San Diego, Tetouan)	M-H	L
[23]	STLF	ISSAE + edRVFL	Commercial building data	M-H	M
[24]	STLF	CEEMDAN + BiLSTM	4 different data sets (UCI, HERC, Malaysia)	H	M
[25]	HSEO	ANN + MIP	Simulation data (TRNSYS, SAM)	H	L-M
Our Work	STLF	ESN + GRU	UCI	L-M	H

Note: The comparison highlights that high-performance DL models such as CNN-LSTM, CEEMDAN-BiLSTM, and KAN-RNN have limitations such as high computational cost and limited edge compatibility. In contrast, the proposed model (ESN-GRU) optimizes the prediction accuracy-computational cost, providing a more practical solution for resource-constrained edge devices.

of DL and ML-based solutions, EdgeAI-based solutions, and solutions that optimize computational efficiency.

A. DL & ML BASED SOLUTIONS

The interest in DL & ML-based solutions is increasing day by day because they provide higher accuracy than traditional statistical-based solutions. Lian et al. [16] proposed a study for building energy forecasting (BEF) with the Light Gradient-Boosting Machine (LGBM) based method. They tested the proposed method in a large-scale simulation and achieved a high amount of accuracy. Ashraf et al. [17] propose a hybrid approach combining ANN, RF, and XGBoost for OEF. Experiments verify the high performance using simulated data from Autodesk Revit and Green Building Studio. In another study, Lu et al. [18] used a hybrid CNN-LSTM-based model to develop a strategy in phase-change thermal storage systems. In [19] and [20], the authors used BiLSTM, GRU, MLP, and GNN-based models with high performance in complex time-series data. Danish and Grolinger [21] introduced the Kolmogorov-Arnold Recurrent Network (KARN) that combines Kolmogorov-Arnold networks and RNN. KARN outperformed traditional RNN-based models (LSTM) on high-variance datasets in STLF applications. However, memory usage and computational complexity are still issues for EdgeAI environments.

Although it has high prediction performance, it is difficult to apply these methods on edge devices. Because such models depend on large data sets and require high processing power.

B. EDGEAI BASED SOLUTIONS

This subsection examines energy estimation systems applicable to EdgeAI environments. In [22], the authors propose an edge computing-based system (Raspberry Pi 4b) for STLF applications using Inverted Transformer (iTx) and Transfer Learning (TL). Thus, they present an EdgeAI framework with low power consumption by providing a low latency (edge computing) and minimal data learning (TL). Liu et al. [23] estimate the occupancy rate of a building using wireless sensor data. Thus, the building electricity load is estimated using the occupancy information of the building with

ISSAE + edRVFL based models. Although this model has high potential to be integrated into the EdgeAI environment, its computational cost is high.

C. COMPUTATION OPTIMIZE SOLUTIONS

DL-based models generally show high performance on large volumes of data, which also means high CPU and memory consumption during model training. To solve this problem, lightweight DL-based models CEEMDAN and Residual BiLSTM are used to achieve high prediction performance and low computational cost [24]. Although the results show higher prediction performance compared to traditional DL models (LSTM and CNN), their integration into the EdgeAI environment requires additional optimization. Tang et al. [25] proposed a hybrid model combining ANN + MIP for hybrid storage energy optimization. Experiments show positive improvements in energy consumption and electricity cost. However, complex optimization techniques are not suitable for EdgeAI environments because they increase computational cost.

Critical Analysis: Table 1 shows the comparative analysis of the reviewed studies and the proposed studies in terms of model type, dataset size, computational cost and edge compatibility.

Considering the Computational Cost and Edge Compatibility, CNN-LSTM + PSO, CEEMDAN + BiLSTM, and KAN + RNN hybrid approaches have low edge compatibility since they will require high processing power and computational cost. LGBM and ANN-based approaches are expected to have lower prediction performance despite having lower computational cost.

Models designed to be compatible with EdgeAI, such as ISSAE + edRVFL and iTx + TL try to balance medium-high computational efficiency and prediction performance, but their EdgeAI compatibility is low (Transfer Learning (TL) Load + Transformer architecture). On the other hand, ESN-GRU emerges as a more reasonable option with its low to medium computational cost and high edge compatibility. Fast inference and low RAM consumption make it stand out as an efficient solution in STLF.

III. METHODOLOGY

This section describes a new model, the EdgeAI-Driven Hybrid ESN-GRU model with low latency and high accuracy for STLF applications. In the first subsection, the relationship between demand response systems and smart grids is explained, and why accurate load forecasting is important is explained. In the second and third subsections, the ESN-GRU model and its mathematical background are given. In the fourth subsection, the research organization is given so that the reader can better understand the flow of the paper. In the fifth and sixth Subsections, the dataset, and performance metrics used in the performance comparison are given, and the section is concluded.

A. THE RELATION OF SMART GRIDS AND DEMAND RESPONSE SYSTEMS

Smart grids are a new generation of power systems that establish an efficient and sustainable balance between energy production and consumption. With the integration of RESs into these systems (smart grids), ensuring supply-demand balance is of critical importance. Figure 2 depicts smart grid integration with RESs such as wind turbines and solar panels. Energy from RESs is transmitted to the grid via battery storage systems, while energy generated from traditional sources such as coal/nuclear power plants contributes to these grids. Smart Meters collect consumers' energy consumption data in real time and thus contribute to the production-consumption balance. Demand Response Systems inform consumers when energy consumption is intense and more expensive.

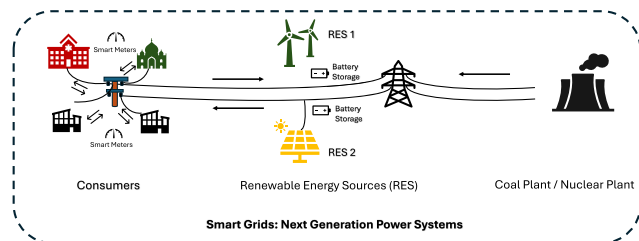


FIGURE 2. Integration of renewable energy sources (RES) into the smart grid.

B. EDGEAI-DRIVEN HYBRID ESN-GRU MODEL

Smart grids have variable structures in terms of energy management and stability due to their heterogeneous structure (traditional energy sources and RES sources), and these problems can be reduced with STLF. STLF usually requires DL models that require high processing power, and the integration of these models on low-power edge devices is a significant challenge. This paper proposes the EdgeAI-supported Hybrid ESN-GRU model to solve this problem. The ESN-GRU model combines the strengths of two different architectures: (i) the Echo State Network (ESN) with fast training time and low computational capabilities, and (ii) the Gated Recurrent Unit (GRU) models with

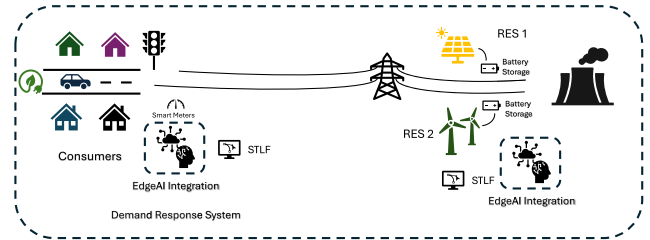


FIGURE 3. Integration of the proposed ESN-GRU model within the smart grid.

high performance in modeling historical dependencies in time series. Thus, ESN-GRU makes low-latency and high-accuracy predictions.

Figure 3 shows the integration of the EdgeAI-supported Hybrid ESN-GRU model into the smart grid infrastructure. Consumer data measured by smart meters is processed on an edge device (Raspberry Pi 4) to perform real-time STLF predictions. Since the data is processed close to the data sources instead of a centralized cloud, low latency and low bandwidth advantages are provided. In this way, the system can make faster and on-site decisions. Contrary to the general smart grid view given in Figure 2, this figure highlights how the ESN-GRU model provides low-latency predictions by reducing dependency on centralized cloud resources.

Figure 4 details the general architecture of ESN-GRU for STLF. The first step involves collecting real-time data from residential, vehicular, and industrial sources in smart grids by sensors. This data is then preprocessed (e.g., normalization) for efficient training. The normalized data is fed into the hybrid database, which includes ESN for fast temporal feature transformation and GRU for learning sequential dependencies. The hybrid model is deployed on the edge device (Raspberry Pi) and performs predictions locally, without relying on cloud resources. As a result, ESN-GRU enables low-latency predictions, guaranteeing low bandwidth consumption even in offline scenarios.

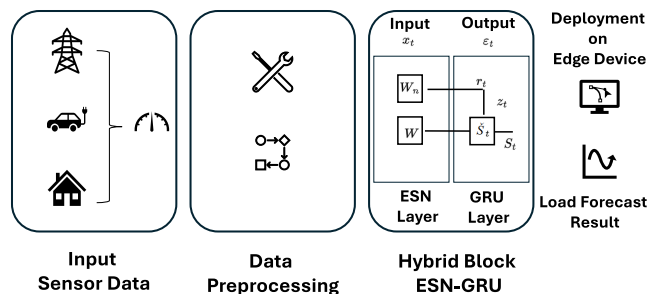


FIGURE 4. Overall architecture of the proposed ESN-GRU model for short-term load forecasting.

Figure 5 shows the real hardware setup where the ESN-GRU model is deployed in this paper. Raspberry Pi 4 is used as the edge device and the model deployed on the device will be shown to be low-latency and energy efficient through experiments in section IV.



FIGURE 5. Experimental EdgeAI deployment environment for STLf.

Algorithm 1 describes the architecture of the pseudo code of ESN-GRU. In the first step (1), each feature is normalized to have a mean of zero and a standard deviation of one. This is done to prevent dominance among features and to achieve faster convergence. $x_t^{(i)}$ denotes the value of the i th feature at time t . μ and σ are the mean and standard deviation, respectively. In the second step (2), the reservoir, which is the fundamental component of ESN, is created to keep it stable. W is the reservoir weight matrix, the spectral radius ρ , and $\lambda_{\max}(W)$ is the largest eigenvalue of W . In the third step, the internal states of ESN over time are calculated. Here α represents the reservoir update rate, h_t represents the reservoir state at time t , and these vectors are given as input to the GRU layer. The fourth and eighth rows show the training of the GRU layer. In the last step, the predicted values (y_t) are returned to the original scale by the denormalization process.

C. MATHEMATICAL BACKGROUND OF ESN-GRU MODEL

The model proposed in this paper, ESN-GRU, combines the unique temporal feature-capturing capabilities of the Echo State Network (ESN) dynamic reservoir and the Gated Recurrent Unit (GRU). Thus, it works well in edge environments with low computational complexity and high temporal prediction performance. In this subsection, the mathematical structure of ESN-GRU is explained.

Echo State Network (ESN):

The ESN model includes three main layers: input weights (W_{in}), reservoir weights (W), and output weights (W_{out}).

Algorithm 1 Pseudo Code of ESN-GRU

Input: $X_{train}, Y_{train}, X_{test}, \mathcal{P}_{ESN-GRU}$

Output: $\hat{Y}_{test}$

Data Normalization

$$1: \hat{x}_t^{(i)} = \frac{x_t^{(i)} - \mu_t^{(i)}}{\sigma^{(i)}}$$

Creation of a Reservoir (ESN)

$$2: W \leftarrow \frac{W}{\lambda_{\max}(W)} \cdot \rho$$

Calculation of Internal Situations (ESN)

$$3: h_t = (1 - \alpha) \cdot h_{t-1} + \alpha \cdot \tanh(W_{in}x_t + Wh_{t-1}), \quad h_0 = 0$$

Training of the GRU Layer

$$4: z_t = \sigma(W_z h_t + U_z s_{t-1} + b_z)$$

$$5: r_t = \sigma(W_r h_t + U_r s_{t-1} + b_r)$$

$$6: \tilde{s}_t = \tanh(W_h h_t + U_h(r_t \odot s_{t-1}) + b_h)$$

$$7: s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \tilde{s}_t$$

$$8: \hat{y}_t = W_o s_t + b_o$$

$$9: \hat{y}_t^{(i)} = \hat{y}_t^{(i)} \cdot \sigma^{(i)} + \mu^{(i)}$$

10: **End**

The mathematical structure for all these components can be explained as follows.

$$h_t = (1 - \alpha)h_{t-1} + \alpha \cdot \tanh(W_{in}x_t + Wh_{t-1}) \quad (1)$$

Here h_t represents the reservoir state, α , leaking rate, x_t , input vector for time period t .

Gated Recurrent Unit (GRU):

The GRU model controls the information flow using update and reset gates (via update and reset gates (h_t)).

$$z_t = \sigma(W_z h_t + U_z s_{t-1} + b_z) \quad (2)$$

$$r_t = \sigma(W_r h_t + U_r s_{t-1} + b_r) \quad (3)$$

$$\tilde{s}_t = \tanh(W_h h_t + U_h(r_t \odot s_{t-1}) + b_h) \quad (4)$$

$$s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \tilde{s}_t \quad (5)$$

Here, z_t represents gate update, r_t represents gate reset, $\tilde{s}_t$ represents candidate state, s_t represents final state, $\odot$ represents multiplication, $\sigma(\cdot)$ represents sigmoid activation function.

Output Layer: For the output, the GRU hidden state is passed through a linear layer, and denormalization is applied for the original scaled values:

$$\hat{y}_t = W_o s_t + b_o \quad (6)$$

$$\hat{y}_t^{(i)} = \hat{y}_t^{(i)} \cdot \sigma^{(i)} + \mu^{(i)} \quad (7)$$

The ESN-GRU hybrid approach thus captures short-term dynamics and long-term dependencies and offers advantages such as low latency in environments such as edge devices.

Computational Complexity Analysis: When calculating the computational complexity for the ESN-GRU model, particular attention should be paid to the combination of the ESN reservoir update and the GRU sequential operations. In the ESN part, the reservoir update requires $\mathcal{O}(N_r \cdot d + N_r^2)$ operations, where N_r and d represent the reservoir and input sizes, respectively. In the GRU part, there are $\mathcal{O}(h^2 + h \cdot d)$ operations, and h represents the number of hidden

units. Consequently, the overall complexity of ESN-GRU is $\mathcal{O}(N_r^2 + h^2)$.

The baseline CEEMDAN-BiLSTM has a complexity of $\mathcal{O}(T \cdot h^2)$ per sequence and involves expensive signal decomposition and bidirectional iteration. The complexity for transformer-based iTx+TL is $\mathcal{O}(T^2 \cdot d)$. Consequently, ESN-GRU has lower theoretical complexity, which translates to faster inference and less energy consumption on edge devices. The results are consistent with computational complexity analysis.

D. DATASET

This study uses the “House4.csv” dataset, part of the public Residential Electricity Consumption of Pakistan (PRECON) database. The PRECON dataset was created by the Lahore University of Management Sciences (LUMS) by collecting electricity consumption data at minute resolution from 42 households in Lahore, Pakistan. For each household, the data includes whole-house load measurement, appliance-level monitoring (e.g., air conditioners, refrigerators, washing machines, kitchen appliances), and demographic information such as number of occupants, property size, and electrical connection type. This dataset was selected because it is one of the most comprehensive residential energy datasets in South Asia and provides detailed insights into real-world smart grid and IoT environments.

TABLE 2. Main abbreviations in the dataset and their explanations.

Abbreviation	Explanation
HVAC	Heating, Ventilation, and Air Conditioning
REFR	Refrigerator
WM	Washing Machine
KA	Kitchen Appliances (e.g., oven, microwave)
BR_kW	Bedroom Load
ACBR_kW	Air Conditioning Load in the Bedroom
Usage_kW	Total Active Power Consumption of the House (kW)

The “House4.csv” dataset used in this paper contains detailed appliance-level consumption profiles for a single household in PRECON, with monthly, weekly, and hourly records. Figure 6 shows the total energy consumption of a house in a week and reveals clear cyclical patterns depending on the usage of appliances. The ESN-GRU model can provide important information by effectively capturing these temporal fluctuations. Figure 7 shows the average power consumption for different appliances. Looking at the graph, it can be seen that the total consumption of HVAC and kitchen appliances is higher than other appliances, and insights can be provided regarding high-efficiency appliances. Figure 8 shows the share (in percentages) of these devices in the total home energy consumption. It provides important information in scenarios such as EdgeAI, where the input size needs to be optimized to reduce the computational burden. Although the dataset represents a single household, its detailed structure at the appliance level allows for realistic short-term load changes when evaluating EdgeAI-based prediction models.

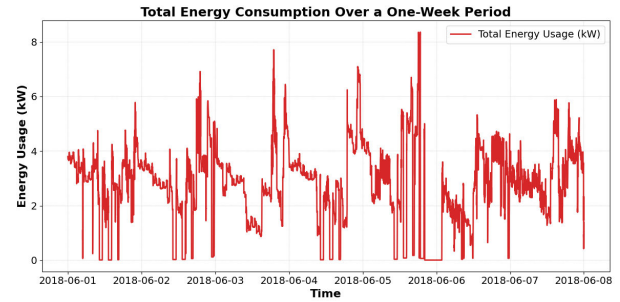


FIGURE 6. Total energy consumption over a one-week period.

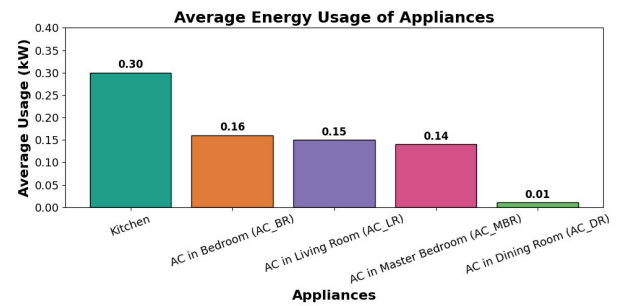


FIGURE 7. Average energy usage of different appliances.

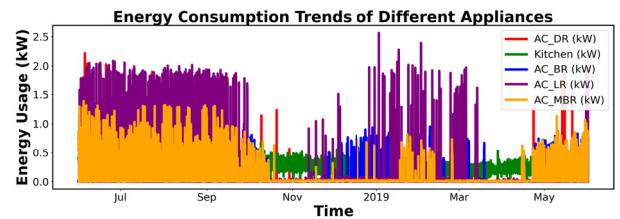


FIGURE 8. Energy usage distribution among different household appliances.

Furthermore, the ESN-GRU model, with its lightweight architecture, can be trained with fewer parameters and can be trained reliably with this data scale. Also, the abbreviations in the paper are given in Table 2.

E. EVALUATION CRITERIA

This subsection introduces the measurement metrics that will be used throughout the paper, especially in Section IV, to evaluate the predictive performance and computational efficiency of the models.

Mean Absolute Percentage Error (MAPE): This metric shows the average of the percentage errors between the predicted ($\hat{y}_t$) and the true values (y_t). Therefore, it means that among the models examined, the model with the lowest MAPE shows better prediction performance.

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (8)$$

Mean Absolute Error (MAE): This metric shows the average of absolute deviations between the predicted ($\hat{y}_t$) and the true values (y_t). Therefore, it means that among the

models examined, the model with lower MAE shows better prediction performance.

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (9)$$

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): These metrics are used to evaluate the prediction variance of the models and penalize large errors to a large extent. Therefore, in the studied models, the model with lower MSE and RMSE shows better prediction performance.

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (11)$$

R-squared Score (R^2): This metric shows the proportion of variance explained by the models. Models with R^2 values closer to 1 indicate better performance.

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2} \quad (12)$$

Computation Time Metrics: For all models, training and prediction times were tested in two different environments (Local PC and Raspberry Pi 4) for distribution efficiency and real-time capacity. The results are presented in Section IV.

- Local PC: Intel i7, 16 GB RAM
- Raspberry Pi 4: ARM Cortex-A72, 4 GB RAM

Energy Consumption Estimation: In order to evaluate the resource efficiency of the models, energy consumption was calculated using the following formula:

$$E = P \times T \quad (13)$$

where E represents energy consumed (in Joules), P represents average power consumption of the device (in Watts), and T represents inference time (in seconds). This metric is important because it gives an idea of the real operating cost and provides important insights in the evaluation for energy-constrained environments (smart buildings or IoT).

Clarification on Error Metrics: While MAE, MSE, and RMSE are frequently used in forecasting models, these metrics capture different aspects of model error. In the experiments conducted in this paper, the ESN-GRU model consistently and significantly lowers MAPE (see section IV) compared to the baseline values, despite slightly higher MAE and RMSE. Because MAPE is a relative percentage-based metric, it is considered more informative in STLF tasks where load values vary. [26]. Despite lower absolute errors in the baseline results, high MAPE values (over 100%) are observed, indicating that these forecasts can cause significant deviations and errors in the actual load profiles. For all these reasons, the ESN-GRU model provides more reliable forecasts in smart grid applications.

TABLE 3. System specifications used in the experiments.

MC: Main Computer, RPI4: Raspberry Pi 4.		
Specification	MC	RPI4
CPU	Intel® Core™ i7-10750H	Broadcom BCM2711, Quad-core Cortex-A72
Clock Speed	12M Cache, up to 5.00 GHz	1.5 GHz Quad-Core
RAM	16 GB	4 GB LPDDR4
OS	Windows 10 Pro	Raspberry Pi OS (Linux)
TDP (W)	45	7.6

IV. PERFORMANCE EVALUATION

This section describes the experimental setup, the data loads used, and the compared base models before moving on to the performance evaluation for the ESN-GRU model. Then, the section concludes by interpreting the findings from the experiments.

A. EXPERIMENTAL SETUP

Experiments were conducted on two different environments to test the performance of the proposed model ESN-GRU under different platforms and hardware specifications. The first environment is a computer with high processing power, and the second environment is an edge device (Raspberry Pi 4) with low energy consumption and hardware capacity. The main reason for making a comparison using two hardware can be explained as follows: (i) our aim of using a limited processing power environment (Raspberry Pi) is to show that the proposed model can run efficiently even in constrained environments, (ii) to highlight the scalability of the model with a high-performance computing setup and to materialize the differences in latency, energy consumption and resource usage on heterogeneous platforms.

Table 3 shows the technical specifications of both systems. When looking at the table, the first system (PC) has higher hardware specifications and is used to evaluate the potential performance of the models in high hardware environments. The second system (Raspberry Pi 4) is used to simulate EdgeAI scenarios and to demonstrate the superior flexibility and energy efficiency of the proposed model compared to baselines in resource-constrained environments.

Table 4 shows the hyperparameter values used in the training of all compared models. In order to provide a fair comparison, the training cycle (epoch), optimizer type and learning rate are kept constant in all models. In this way, the model performance is evaluated only due to architectural differences. As a result of the hyperparameter optimization performed while training the ESN-GRU model, the most optimum hidden units value was determined as 256, the optimization algorithm was Adam, and the learning rate was 0.001. Similarly, the most optimum values were determined as a result of the hyperparameter optimization for the other baselines.

B. WORKLOADS

In order to evaluate all models in real-world conditions, the publicly published “House4.csv” dataset of LUMS Smart

TABLE 4. Hyperparameters of compared models.

A: Architecture, HU: Hidden Units, O: Optimizer LR: Learning Rate E: Epochs

Model	A	HU	O	LR	E
Base 1 [24]	Residual BiLSTM	64	Adam	0.001	50
Base 2 [22]	iTx Transfer Learning	64 (key_dim)	Adam	0.001	50
ESN-GRU	Hybrid ESN GRU	256	Adam	0.001	50

Grids Lab is used.¹ This dataset contains hourly energy consumption information of various electrical devices (such as HVAC systems, refrigerators, washing machines) in the house, which has a structure suitable for STLF studies. The main motivation behind choosing this dataset is that it has daily and weekly cyclical energy consumption patterns and that the input dimensions can be simplified for EdgeAI scenarios.

Before training all models, the dataset was subjected to data preprocessing to clean missing values, filter anomalies, and feature normalization. Then, the dataset was divided into 80.

C. BASELINE MODELS

In order to evaluate the performance of the proposed model ESN-GRU for STLF, two new baseline models found in the literature are selected. These models are explained in this subsection.

- **CEEMDAN + BiLSTM Approach [24]:** In this study, the authors propose a new hybrid model that combines Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Bidirectional Long Short-Term Memory (BiLSTM) models. CEEMDAN model decomposes different frequency components in the input signals into intrinsic mode functions (IMFs). And these IMFs are modeled by BiLSTM to learn temporal dependencies in forward and backward directions. Although the model works with high accuracy, its computational complexity and long training time make it less suitable for latency and resource constrained edge environments.
- **iTx + Transfer Learning (TL) Approach [22]:** In this study, the authors proposed a novel hybrid model that combines Inverted Transformer (iTx) and Transfer Learning (TL) models. iTx embeds time series per feature with a transformer architecture, which makes it suitable for multivariate time series forecasting. TL is used to improve the forecasting performance in scenarios where there is little data. Since the model is dependent on transformer architecture and domain adaptation techniques, the complexity of implementation and training time is high.

Lightweight models such as XGBoost and LGBM are frequently used in STLF tasks due to their lower complexity.

¹<https://web.lums.edu.pk/~eig/CXyzsMgyXGpW1sBo>

TABLE 5. Comparison of ESN-GRU with baselines in STLF.

Study	R ²	MAPE	MAE	MSE	RMSE
Base 1 [24]	0.961	129.219	0.013	0.001	0.025
Base 2 [22]	0.954	132.767	0.015	0.001	0.027
ESN-GRU	0.986	0.101	0.097	0.021	0.145

Note: The MAPE values in the table represent percentages (%). Due to large relative errors in the baseline models, the MAPE values exceed 100%. For the proposed model ESN-GRU, this value is approximately 10.1%.

However, we did not include them in this paper because we focused on current DL and EdgeAI-focused methods, and previous studies ([16], [17]) reported lower accuracy compared to hybrid DL approaches on nonlinear and high-variance datasets. For all these reasons, the ESN-GRU hybrid model is expected to perform better (compare to Lightweight ML models) in edge scenarios.

D. RESULTS

In this subsection, the ESN-GRU model and baselines are comparatively evaluated with respect to STLF performance and all experiments are performed using a high-end computer and an edge device with limited resources, Raspberry Pi 4. First, STLF accuracy is evaluated, then computational efficiency parameters are evaluated and EdgeAI compatibility of the models is interpreted and the section is concluded.

STLF Performance Evaluation: Table 5 shows the STLF performance results for all models. The ESN-GRU model outperformed the baselines in all performance metrics. When the results are carefully examined, it is seen that the proposed ESN-GRU model shows 2.5% higher variance with an R^2 score of 0.986 compared to the most successful baseline model, and its MAPE value is 0.101, providing a significant improvement over the other baseline models. For the ESN-GRU model, the MAE value of 0.097 and the RMSE value of 0.145 indicate high prediction accuracy.

Discussion on MAE, MSE, and RMSE: The results in Table 5 confirm that the ESN-GRU model has a lower MAPE (0.101) and significantly higher performance compared to the results of the (proposed) baseline models (129 and 132). In contrast, the MAE (0.097) and RMSE (0.145) values for ESN-GRU are slightly higher than those of the baseline models, but this difference is small in absolute terms. This apparent discrepancy stems from the fact that the baseline models have artificially small absolute errors in certain segments, but cannot generalize to the general load pattern. This is confirmed by the very high MAPE values of the baseline models. ESN-GRU balances both relative and absolute error metrics and produces consistent results. A careful examination of Figure 9 reveals that ESN-GRU converges faster and has a more stable training behavior in the RMSE curve. As a result, ESN-GRU produces more consistent results for real-world smart grid operations by predicting sudden short-term fluctuations more accurately than other models.

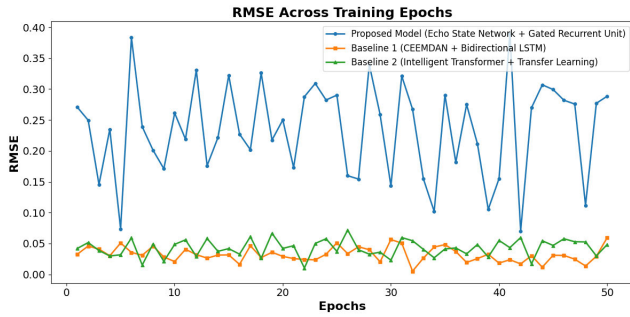


FIGURE 9. RMSE values over different epochs for the proposed ESN-GRU model and baseline models.

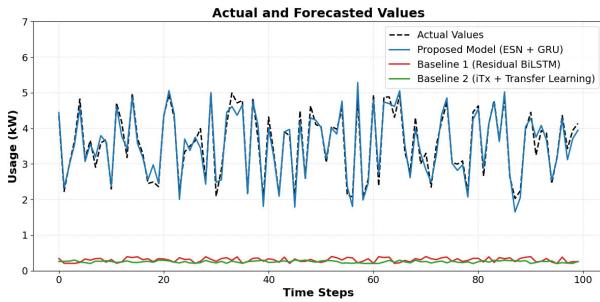


FIGURE 10. Comparison of actual and forecasted values for the proposed ESN-GRU model and baseline models.

Figure 9 shows the RMSE values for all models against varying epoch numbers. When the figure is examined, it is seen that the ESN-GRU model converges more stably and quickly. This shows that ESN-GRU reaches lower error levels faster and remains stable. On the other hand, the baseline models (CEEMDAN + BiLSTM and iTx + TL) converge later and show more fluctuations during the training. This shows that the ESN-GRU model has higher accuracy and a more stable training process compared to the other models.

Figure 10 presents the visualization of the distribution of the forecast results of all models with respect to the actual value. The results show that the ESN-GRU model can learn short-term changes and temporal dependencies much better than other models. In particular, it can be seen that the model is quite close to the actual consumption curve at the points where sudden changes occur. This shows the model's sensitivity to short-term fluctuations and its capacity to learn structural patterns in time series.

Computational Efficiency and EdgeAI Compatibility:

Table 6 shows the computational efficiency (training & prediction times) of all models on both platforms. The results show that the ESN-GRU model has the lowest training and prediction time on all platforms and is more suitable for edge environments, especially with a performance time of only 1.56 seconds on Raspberry Pi 4. The comparison performed using these two different platforms demonstrates the deployability of ESN-GRU in both traditional computing environments and edge-based IoT infrastructures. It can operate efficiently even in constrained environments, while also

TABLE 6. Comparison of model training and prediction times on two different platforms.

Model	PC	
	Training (ms)	Prediction (ms)
[24]	675,239	997
[22]	166,630	383
ESN-GRU	160,926	79
Model	RPi 4	
	Training (min)	Prediction (sec)
[24]	222.12	19.68
[22]	54.81	7.56
ESN-GRU	52.94	1.56

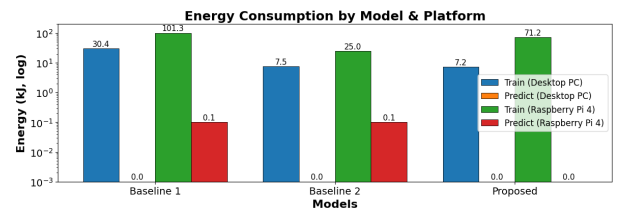


FIGURE 11. The energy consumption comparison of models and platforms according to training and prediction time.

demonstrating the model's scalability in high-performance environments and the differences in latency, energy consumption, and resource utilization across heterogeneous platforms.

Figure 11 shows the energy consumption of all models for training and prediction processes on both platforms. The energy consumptions are obtained by multiplying the device power requirement (Watt) by the processing time (second/minute) using equation 13 explained earlier. This metric is especially important in energy and performance-constrained environments such as edge devices (EdgeAI). The results prove that the ESN-GRU model has the lowest energy consumption on all platforms.

As a result, the experiments confirm that the ESN-GRU model has lower computational efficiency compared to other baselines and therefore has higher compatibility for EdgeAI environments.

V. CONCLUSION AND FUTURE WORK

Short-Term Load Forecasting (STLF) is the short-term forecasting of electricity demand in energy systems and is frequently used to balance the supply and demand of variable RES resources (such as solar and wind energy). Despite this potential, STLF still brings high computational cost, dependency on cloud infrastructure, and latency issues. To solve these issues, this paper introduces a new Hybrid Echo State Network and Gated Recurrent Unit (ESN-GRU) model based on EdgeAI. While EdgeAI offers the advantages of low latency and low bandwidth by bringing the processing power closer to the data source, ESN-GRU provides advantages such as fast training and low

complexity for edge environments. ESN-GRU and state-of-the-art baselines are compared using real-world data. The results confirm that ESN-GRU achieves a significant drop in MAPE (from 129 to 0.101) and a 2.5% increase in R^2 compared to other baselines. Additionally, the ESN-GRU model demonstrated up to 92% faster inference time and significantly lower power consumption for EdgeAI environments such as Raspberry Pi 4.

There are some aspects of the ESN-GRU model that need to be explored and developed in future studies:

- **Model Generalization for Different Environments:** Only one subject dataset is used while evaluating the current model. Therefore, in order to generalize the model better, it can be evaluated using data with different types of buildings and varying energy consumption patterns.
- **Multi-step Prediction:** While one-step prediction is made in the current model, multi-step predictions can be made for demand response and energy scenarios in future studies.
- **Federative Learning Integration:** Federative learning mechanisms can be explored to protect the privacy of consumers in smart grids.
- **Practical Deployment Scenarios:** In practice, the ESN-GRU model can be integrated into demand response programs where energy distribution companies notify consumers of peak demand periods as a result of STLF. Furthermore, ESN-GRU can be deployed in IoT hubs to predict device-based consumption in smart home platforms. For example, it can optimize the scheduling of high-power devices like HVAC systems.
- **Error Metrics Optimization:** Additionally, although the ESN-GRU model has better MAPE and R^2 than the baseline models, future studies could reduce the MAE and RMSE values. This could be achieved by applying hyperparameter tuning (e.g., reservoir size, hidden units) and conducting experiments on larger multi-digit datasets.

SOFTWARE AVAILABILITY

The dataset is publicly published for future researchers at: <https://github.com/NextGen-Wireless-Security-Group/ESN-GRU>.

REFERENCES

- [1] A. Kumar, M. Alaraj, M. Rizwan, and U. Nangia, "Novel AI based energy management system for smart grid with RES integration," *IEEE Access*, vol. 9, pp. 162530–162542, 2021.
- [2] A. S. Brouwer, M. Van Den Broek, A. Seebregts, and A. Faaij, "Impacts of large-scale intermittent renewable energy sources on electricity systems, and how these can be modeled," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 443–466, May 2014.
- [3] P. Zhang, X. Wu, X. Wang, and S. Bi, "Short-term load forecasting based on big data technologies," *CSEE J. Power Energy Syst.*, vol. 1, no. 3, pp. 59–67, Sep. 2015.
- [4] K. N. Singh and K. R. Singh, "A review on deep learning models for short-term load forecasting," in *Proc. Appl. Artif. Intell. Mach. Learn.*, 2021, pp. 705–721.
- [5] F. Bayram, P. Aupke, B. S. Ahmed, A. Kassler, A. Theocharis, and J. Forsman, "DA-LSTM: A dynamic drift-adaptive learning framework for interval load forecasting with LSTM networks," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106480.
- [6] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud computing: An overview," in *Proc. IEEE Int. Conf. Cloud Comput.*, Apr. 2009, pp. 626–631.
- [7] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D. S. Nikolopoulos, "Challenges and opportunities in edge computing," in *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, Nov. 2016, pp. 20–26.
- [8] W. Holderbaum, F. Alasali, and A. Sinha, "Short term load forecasting (STLF)," in *Energy Forecasting and Control Methods for Energy Storage Systems in Distribution Networks: Predictive Modelling and Control Techniques*. Cham, Switzerland: Springer, 2023, pp. 13–56.
- [9] K. Metaxiotis, A. Kagiannas, D. Askounis, and J. Psarras, "Artificial intelligence in short term electric load forecasting: A state-of-the-art survey for the researcher," *Energy Convers. Manage.*, vol. 44, no. 9, pp. 1525–1534, Jun. 2003.
- [10] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proc. IEEE*, vol. 111, no. 1, pp. 42–91, Jan. 2023.
- [11] S. B. Melhem, M. Golec, A. Alwarafy, and Y. Khamayseh, "LENS: Lightweight and explainable LLM-based APT detection at the edge for 6G security," *IEEE Access*, vol. 13, pp. 172402–172415, 2025.
- [12] Y. Wang, C. Yang, S. Lan, L. Zhu, and Y. Zhang, "End-edge-cloud collaborative computing for deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2647–2683, 4th Quart., 2024.
- [13] A.-X. Wang and J.-J. Li, "A novel cloud-edge collaboration based short-term load forecasting method for smart grid," *Frontiers Energy Res.*, vol. 10, Aug. 2022, Art. no. 977026.
- [14] H. B. Gooi, T. Wang, and Y. Tang, "Edge intelligence for smart grid: A survey on application potentials," *CSEE J. Power Energy Syst.*, vol. 9, no. 5, pp. 1623–1640, 2023.
- [15] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.
- [16] H. Lian, Y. Ji, M. Niu, J. Gu, J. Xie, and J. Liu, "A hybrid load prediction method of office buildings based on physical simulation database and LightGBM algorithm," *Appl. Energy*, vol. 377, Jan. 2025, Art. no. 124620.
- [17] J. Ashraf, R. Azam, A. A. Rifa, and M. J. Rana, "Multiple machine learning models for predicting annual energy consumption and demand of office buildings in subtropical monsoon climate," *Asian J. Civil Eng.*, vol. 26, no. 1, pp. 293–309, Jan. 2025.
- [18] S. Lu, H. Wei, Y. Jia, R. Wang, Y. Sun, and Q. Yang, "A control method of electric boiler phase change thermal storage heating system based on dual-time scale load prediction model," *J. Energy Storage*, vol. 107, Jan. 2025, Art. no. 114959.
- [19] R. Akter, M. G. Shirkoobi, J. Wang, and W. Mérida, "An efficient hybrid deep neural network model for multi-horizon forecasting of power loads in academic buildings," *Energy Buildings*, vol. 329, Feb. 2025, Art. no. 115217.
- [20] X. Dong, Y. Luo, S. Yuan, Z. Tian, L. Zhang, X. Wu, and B. Liu, "Building electricity load forecasting based on spatiotemporal correlation and electricity consumption behavior information," *Appl. Energy*, vol. 377, Jan. 2025, Art. no. 124580.
- [21] M. U. Danish and K. Grolinger, "Kolmogorov-arnold recurrent network for short term load forecasting across diverse consumers," *Energy Rep.*, vol. 13, pp. 713–727, Jan. 2025.
- [22] M. S. Iqbal and M. Adnan, "Edge computing and transfer learning-based short-term load forecasting for residential and commercial buildings," *Energy Buildings*, vol. 329, Feb. 2025, Art. no. 115273.
- [23] C. Liu, Z. Xu, M. Yuan, J. Xie, Y. Yuan, and K. Ma, "Building electrical load forecasting with occupancy data based on wireless sensing," *Appl. Energy*, vol. 380, Feb. 2025, Art. no. 124960.
- [24] J. Han and P. Zeng, "Residual BiLSTM based hybrid model for short-term load forecasting in buildings," *J. Building Eng.*, vol. 99, Apr. 2025, Art. no. 111593.
- [25] H. Tang, B. Li, Y. Zhang, J. Pan, and S. Wang, "A coordinated predictive scheduling and real-time adaptive control for integrated building energy systems with hybrid storage and rooftop PV," *Renew. Energy*, vol. 239, Feb. 2025, Art. no. 122047.

- [26] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H.-A. Jacobsen, "Household electricity demand forecasting: Benchmarking state-of-the-art methods," in *Proc. 5th Int. Conf. Future energy Syst.*, Jun. 2014, pp. 233–234.



SUHIB BANI MELHEM received the M.Eng. and Ph.D. degrees in electrical and computer engineering from Concordia University, Canada. He completed a postdoctoral fellowship at York University in collaboration with the National Research Council Canada. In 2023, he earned a Certificate in Cybersecurity from Polytechnique Montréal. From May to August 2023, he was a Scientific Researcher of cybersecurity with Zayed University, United Arab Emirates, where he focused on integrating cutting-edge advancements in cybersecurity. He is currently an Assistant Professor with Al Ain University. His research interests include cloud computing, cybersecurity, information security, machine learning, the IoT, and 5G networks.



MUHAMMED GOLEC received the M.Sc. and Ph.D. degrees in computer science from the Queen Mary University of London, under the prestigious Turkish Ministry of National Education Scholarship. He is currently an Assistant Professor with the Department of Computer Engineering, Bogaziçi University. During his post-graduate studies, he authored more than 40 peer-reviewed publications in high-impact journals and leading international conferences, including IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, *ACM Computing Surveys*, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE CCGrid. He gained industry experience by participating in Industry 4.0 projects as an Electrical and Electronic Maintenance Engineer at Sisecam, a global glass manufacturing company that combines theoretical knowledge with applied industrial expertise. His research interests include artificial intelligence, cloud computing, and security and privacy. He has also received multiple awards in recognition of his research and reviewing contributions: Demonstrator of the Year (2024 Queen Mary University of London), Reviewer Awards (2025 Wiley TETT, 2024 Elsevier IoT, and 2023 Wiley TETT), and Top Cited Article (2022–2023 Wiley SPJ).



SAED ALRABAAEE received the Ph.D. degree in information system engineering from Concordia University, Montreal, Canada. He is currently an Associate Professor with the Department of Information Systems and Security, UAEU. Prior to joining UAEU, he was a Visiting Assistant Professor with the Department of Electrical and Computer Engineering and Computer Science, University of New Haven (UNH), USA. He is also a permanent Research Scientist with the Security Research Center, CIISE, Concordia University. His research and development activities and interests focus on the broad area of reverse engineering, including binary authorship attribution and characterization and malware investigation. In this domain, he has published more than 30 articles in top-tier journals and in prestigious conferences.



MUATH ALSHAIKH received the B.Sc. degree in computer science from Al-Balqa University, Jordan, in 2006, the master's degree in computer science from Uta University, Malaysia, in 2010, and the Ph.D. degree in computer science from the Université de Bretagne Occidentale, France, in 2016. He was an Assistant Professor with the College of Computing and Informatics, Saudi Electronic University, Saudi Arabia. He is currently an Assistant Professor with Al Ain University. He is with the Lab-STICC/UMR CNRS 6283, SFIIS Team, Université de Bretagne Occidentale. His research interests include image and video watermarking, cryptology, information security, image processing, the Internet of Things (IoT), cybersecurity, and computer vision.



MURAT UYAR received the M.Sc. degree in electrical sciences and the Ph.D. degree in electrical and electronics engineering from Firat University, Elâzğ, Türkiye, in 2004 and 2008, respectively. He is currently an Associate Professor with the Faculty of Engineering, Department of Electrical and Electronics Engineering, Bursa Uludağ University. His research interests include power systems engineering, power quality, power system control, power electronics, signal processing, and pattern recognition. He has published research articles in high-impact SCI journals. He is actively working on developing mid-frequency resistance spot welding systems, high precision welding applications, and adaptive control of welding systems.

...