An Innovative Approach to Short-Term Load Forecasting Using Sequential Modeling and Advanced Feature Engineering

ECE 666 Distribution System Engineering

Dafinny thanigaivel, Electrical and Computer Engineering, dthaniga@uwaterloo.ca

Abstract— **This project aims to replicate and expand upon the findings presented in *"A Novel Technique for Short-Term Load Forecasting Using Sequential Models and Feature Engineering"*, [6], which proposed a hybrid framework combining sequential models (LSTM, GRU, RNN) with engineered features and timestamp-based training for short-term load forecasting (STLF). In the original study, the DeepDeFF-BGRU model was found to be the best performer, achieving a MAPE of 0.99%, outperforming traditional unidirectional and bidirectional models.**

In this replication, our work evaluates six sequential models—RNN, LSTM, GRU, and their bidirectional variants—within the DeepDeFF architecture across three real-world datasets: ERCOT(The Electricity Reliability Council of Texas), PRECON(Pakistan Residential Electricity Consumption), and RTE(Réseau de Transport d'Électricité). Models are trained and tested across multiple forecast horizons (2, 6, and 12 timesteps) and evaluated using MAPE(Mean Absolute Percentage Error). The results confirm the superior performance of bidirectional and DeepDeFF-based models, particularly DeepDeFF-BRNN and DeepDeFF-GRU, in delivering consistent accuracy across varying data complexities. DeepDeFF-BRNN achieved a MAPE of 4.52% on the ERCOT dataset at the 12-timestep horizon, while DeepDeFF-GRU maintained consistently low MAPE on RTE, reaching 2.54% at 2 timesteps. For the noisier PRECON dataset, DeepDeFF-BLSTM performed best with a MAPE of 13.87% at 12 timesteps. While DeepDeFF variants show strong generalization in low-variance datasets like RTE, challenges remain in high-noise data such as PRECON. This work validates the effectiveness of DeepDeFF under diverse load conditions and highlights areas for further enhancement.

Keywords—Short-term load forecasting, DeepDeFF, BGRU, sequential models, timestamp learning, ERCOT, PRECON, RTE, MAPE.

I. Introduction

The rapid evolution of modern power systems driven by renewable energy integration, electric vehicle adoption, and decentralized energy resources has made **short-term load forecasting (STLF)** a critical tool for grid stability, operational planning, and demand-side management [1], [2]. However, the nonlinear, stochastic nature of electrical loads especially at the household or regional level makes accurate forecasting a challenging task [3].

While traditional models like ARIMA(Autoregressive Integrated Moving Average), exponential smoothing, and support vector regression(SVR) have been widely used, they often fail to capture the complex temporal dynamics of modern load profiles [4], [5], [8]. In contrast, **deep learning models**, particularly **sequential architectures** such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs), offer substantial improvements by modeling long-range dependencies and nonlinear behaviors in time-series data [3], [1]. **Bidirectional variants** of these models, including BLSTM, BRNN, and BGRU, further enhance forecasting by learning from both past and future sequences [6].

To address challenges related to limited data availability and generalization, **Wahab et al.** [6] proposed the **Deep Derived Feature Fusion (DeepDeFF)** framework. This approach combines raw load data with **hand-crafted statistical features** and trains a **Y-shaped sequential model** using **timestamp-based training** to improve performance, particularly on smaller datasets. The original study reported a **MAPE as low as 0.99%**, with **DeepDeFF-BGRU** consistently emerging as the best-performing model across diverse datasets from multiple countries.

Despite these advancements, the reliability and generalizability of hybrid deep learning frameworks like DeepDeFF remain underexplored in varying data environments and operational conditions. Questions remain regarding their sensitivity to temporal resolution, data scarcity, and model complexity—especially when deployed across diverse geographic regions and load types. Furthermore, while models like BGRU have shown **state-of-the-art** performance in experimental settings, their comparative behavior in real-world contexts with different load characteristics and sampling strategies warrants further investigation. This study aims to address these gaps by systematically evaluating and benchmarking sequential deep learning models, with a particular focus on the strengths and limitations of feature fusion architectures in short-term load forecasting.

The next section provides a review of recent advancements in short-term load forecasting, with a focus on sequential deep-learning models and feature engineering techniques. Section III describes the DeepDeFF architecture and its core components. Section IV outlines the experimental setup and datasets used. Section V presents and analyses the results, while the final section discusses key insights and proposes improvements to enhance the adaptability of feature-fusion models like DeepDeFF.

II. Literature survey

Accurate short-term load forecasting (STLF) is a foundational requirement for the stability and efficiency of modern power systems. With increasing penetration of renewable energy sources, dynamic load profiles, and consumer-side technologies, traditional forecasting methods have become insufficient. In response, **deep learning models**, especially those using sequential architectures, have emerged as powerful tools capable of capturing complex temporal and nonlinear patterns in electricity demand.

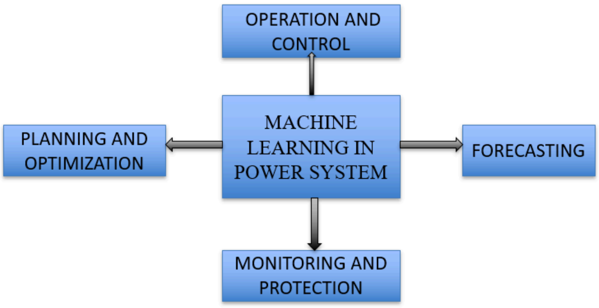
OPERATION AND CONTROL
PLANNING AND OPTIMIZATION
MACHINE LEARNING IN POWER SYSTEM
FORECASTING
MONITORING AND PROTECTION

Fig 1: Machine learning in power system

Machine learning (ML) has emerged as a critical enabler in modern power systems, supporting data-driven solutions across key domains including forecasting, operation and control, planning and optimization, and monitoring and protection. As illustrated in Fig 1, ML enhances the efficiency, reliability, and adaptability of power system operations, addressing challenges posed by increasing complexity, decentralization, and variability in demand and generation.

A. *Classical and machine learning model*

Earlier approaches to STLF relied heavily on statistical models such as ARIMA, linear regression, and exponential smoothing. While these methods are interpretable and fast, they lack the flexibility to capture non-linear dependencies in dynamic load profiles. Machine learning (ML) techniques like decision trees, support vector machines (SVM)[8], and ensemble methods were introduced to improve adaptability and accuracy [1].

Mahajan & Shrivastav [7] conducted an in-depth comparison of regression-based ML models and reported strong performance from **linear SVM** and **bagged tree ensembles**, particularly on datasets with moderate non-linearity. These models also benefited from preprocessing methods like k-fold cross-validation and feature selection.

B. *Emergence of Deep Learning in STLF*

The rise of deep learning has brought a major shift in the field of load forecasting. Unlike classical methods, deep learning models can automatically learn temporal dependencies and nonlinear relationships without extensive feature engineering. Among these, **Recurrent Neural Networks (RNNs)** and their variants—**Long Short-Term Memory (LSTM)** and **Gated Recurrent Units (GRUs)**—are particularly effective in handling time-series data [3]. Kong et al. [3] demonstrated the superiority of LSTM in forecasting residential load data, attributing its success to the model's ability to retain long-term memory through gated connections. Further studies have shown that GRU models offer comparable performance with fewer parameters, making them suitable for applications with limited computational resources.

Yang et al. [4] applied hierarchical deep learning methods to forecast electric vehicle (EV) charging load, proving that deep models can generalize across different energy consumption scenarios when trained on well-preprocessed data.

An enhanced implementation of deep learning for STLF was proposed by **Tsegaye et al. [9]**, who introduced a hybrid model combining **Efficient and Parallel Genetic Algorithms (EPGA)** with LSTM to forecast electricity demand for the Jimma distribution system in Ethiopia. Their model first linearly combines weather and historical load features using EPGA to simplify the LSTM input structure, reducing the complexity and risk of overfitting. The EPGA-enhanced LSTM achieved a **7.486% improvement in RMSE** over vanilla LSTM models. Their work demonstrates that optimized deep models can significantly improve forecasting performance in resource-constrained or highly variable environments—especially in developing countries with emerging power systems.

These developments highlight a growing trend toward not only using deep architectures for STLF but also tailoring them through **domain-specific enhancements** to improve interpretability, accuracy, and scalability. This broader shift from classical statistical approaches to deep, hybrid, and context-aware learning models is summarized in **Fig. 2**, which outlines the evolving landscape of STLF methodologies.

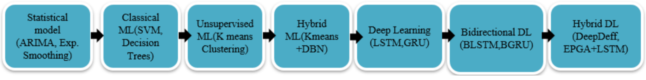
Statistical model (ARIMA, Exp. Smoothing)
Classical ML(SVM, Decision Trees)
Unsupervised ML(K means Clustering)
Hybrid ML(Kmeans +DBN)
Deep Learning (LSTM,GRU)
Bidirectional DL (BLSTM,BGRU)
Hybrid DL (DeepDeff, EPGA+LSTM)

Fig 2: Evolution of short–term load forecasting methods

C. *Bidirectional and hybrid sequential model*

Despite their advantages, traditional RNN-based models only process data in one direction—typically past to future. This restricts the ability of the model to fully learn from surrounding context. **Bidirectional architectures**, including **BLSTM**, **BRNN**, and **BGRU**, address this limitation by training on both past and future data, enabling more context-aware forecasting [7][10].

The **DeepDeFF framework**, proposed by Wahab et al. [6], takes this further by combining bidirectional sequential models with **feature fusion**. It processes raw time-series data and engineered statistical features in parallel using a **Y-shaped architecture**, followed by fusion and prediction layers. A key innovation in DeepDeFF is its

timestamp-based training strategy, where individual models are trained for each time slot in the forecast horizon. This granular approach enhances accuracy but raises challenges around overfitting and computational efficiency.

Zhu et al.'s FR-DBN method [11] also fits within this category, as it blends clustering, deep networks, and optimization techniques into a cohesive hybrid pipeline. By incorporating KMeans and conjugate gradient-enhanced DBNs, their model achieves not only higher accuracy but also significantly reduced training time compared to traditional deep learning methods.

D. DeepDeFF and Timestamp-Based Learning

The **Deep Derived Feature Fusion (DeepDeFF)** framework, proposed by Wahab et al. [6], introduces a hybrid approach to short-term load forecasting by combining **bidirectional sequential models** with **engineered statistical features**. Its dual-input architecture processes raw load sequences in parallel with derived features—such as historical averages and calendar-based indicators—allowing the model to capture both temporal dependencies and domain-informed patterns. A key characteristic of DeepDeFF is its **timestamp-based training strategy**, in which separate models are trained for each prediction interval. This allows for fine-grained learning of time-specific consumption behavior, although it also increases computational requirements and the potential for overfitting in data-scarce environments.

This modeling approach exemplifies the broader shift toward integrating learned and handcrafted features while tailoring training to reflect the structure of temporal demand in power systems.

E. Identified gaps and research

The reviewed literature demonstrates significant advancements in short-term load forecasting (STLF), with a clear evolution from traditional statistical techniques and machine learning approaches to more sophisticated deep and hybrid learning architectures. Despite these developments, several critical challenges remain inadequately addressed across existing studies.

Traditional regression and statistical models [2], while computationally efficient and interpretable, are inherently limited in their ability to capture nonlinear relationships and temporal dependencies in electricity consumption patterns. Although hybrid machine learning frameworks such as **KMeans combined with Feature-Refined Deep Belief Networks (FR-DBN)** [11] have shown improved predictive accuracy by incorporating clustering and hierarchical learning, they often require extensive preprocessing and careful hyperparameter tuning. These factors raise concerns regarding the scalability and standardization of such models in diverse operational environments.

Deep learning models such as LSTM and GRU [5], and their applications to residential[3] and electric vehicle (EV) load forecasting [4], have shown strong sequence learning capabilities. Yet, these models are frequently tested on single datasets under fixed temporal resolutions, limiting the generalizability of their results. Variability in performance across different horizons or geographical contexts is **infrequently addressed in existing studies**, limiting the generalizability of these models.

Architectural optimization strategies, such as the **EPGA-LSTM model** proposed by Tsegaye et al. [9], incorporate efficient feature selection to reduce complexity and improve learning. Nonetheless, their effectiveness remains tightly coupled to the characteristics of the datasets on which they are trained, posing challenges to reproducibility and adaptation in other scenarios.

A recurring limitation across much of the literature is the **lack of cross-dataset benchmarking**. Many models are evaluated in narrowly defined conditions, without consistent experimentation across datasets that vary in resolution, size, or geography. This constrains the ability to assess model robustness under real-world variability. Furthermore, **model interpretability**—an essential factor for deployment in operational power systems—remains insufficiently explored, particularly as forecasting models become increasingly complex and opaque.

In summary, while STLF research has made considerable growth in leveraging deep and hybrid learning models, there remains a pressing need for broader validation, comparative evaluation, and architectural simplification. Future work should prioritize systematic benchmarking, scalable training strategies, and the integration of interpretability tools to support the practical deployment of these models. The present study seeks to address some of these gaps by replicating and assessing the DeepDeFF framework across multiple datasets and forecasting horizons, thereby contributing to a deeper understanding of model generalization, limitations, and operational viability.

III. Proposed methodology

The forecasting framework adopted in this study is designed to evaluate both unidirectional and bidirectional deep learning models across multiple real-world electricity consumption datasets. The overall methodology includes four main components: model architecture selection, sequence generation, feature processing, and hyperparameter configuration.

A. Bidirectional sequential model

To explore the role of directionality in sequence modeling, both unidirectional and bidirectional recurrent neural network (RNN) architectures were employed. The unidirectional models—LSTM, GRU, and Simple RNN—process input sequences from past to future, capturing only historical dependencies. In contrast, bidirectional models (BLSTM, BGRU, BRNN) enhance learning by processing the sequence in both forward and backward directions, allowing the model to access contextual information from the past and the future simultaneously.

Importantly, the bidirectional layer can be constructed using any recurrent cell. In this work, Bidirectional LSTM (BLSTM), Bidirectional GRU (BGRU), and Bidirectional RNN (BRNN) were all implemented to assess performance differences among recurrent cell types within a bidirectional setting.

The general structure of a Bidirectional sequential model, Fig 3 shows the example with LSTM cells (BLSTM), but these cells can be any of the RNN or GRU types[6]. Each input X_t is processed simultaneously by a forward and backward L STM unit. The outputs from both directions are merged and used to produce the final forecast.

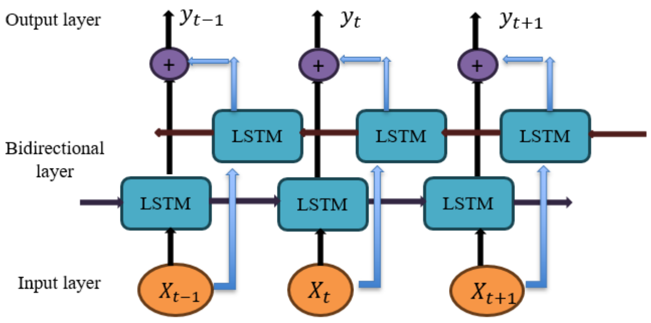

Fig 3: Bi-directional sequence models.

The internal dynamics of a standard LSTM unit, which underpins both forward and backward passes in the BLSTM, are given by [12]:

$$U = F[W_{(t-1)} \ (X_{t-1} + X_t \ ) + H_{t-1} \qquad [1]$$

$$Y = F[W_t \ (X_t - 1 + X_t \ ) + H_t \qquad [2]$$

$$U = F[W_t \ (X_t + X_{t+1} \ ) + H_t \qquad [3]$$

$$Y = F[W_{t+1} \ (X_t + X_{t+1} \ ) + H_{t+1} \qquad [4]$$

Here X_t denotes the input at time step t, Ht represents the hidden state at time t, W_t are the trainable weights at time step t and F(·) is a non-linear activation function (e.g., tanh, ReLU). These expressions abstractly capture the way a bidirectional model utilizes information from neighboring inputs (X_{t-1},X_t, X_{t+1}) and hidden states (H_{t-1},H_t,H_{t+1}) to compute intermediate updates (U) and outputs (Y) at each time step. Similar gating mechanisms apply for GRU cells, albeit with a reduced architecture, and standard RNNs rely solely on hidden state updates without gating.

B. DeepDeFF

In addition to conventional sequential models, this study incorporates the **DeepDeFF (Deep Feature Fusion and Forecasting)** architecture to benchmark advanced temporal learning approaches. DeepDeFF enhances standard recurrent models by combining sequential layers with feature fusion mechanisms that process both raw input features and engineered timestamp-based information. The architecture supports flexibility in temporal learning by allowing any recurrent cell to serve as the backbone.

For implementation, **six DeepDeFF variants** were developed by embedding each of the core recurrent models—**LSTM, GRU, RNN**—and their **bidirectional counterparts (BLSTM, BGRU, BRNN)** into the DeepDeFF framework. This comprehensive setup enables direct comparison between traditional unidirectional, bidirectional, and deep feature fusion approaches. By including DeepDeFF in the experimental pipeline, this work provides a holistic evaluation of sequential modeling strategies for short-term load forecasting across diverse datasets and temporal resolutions.

C. Converting data into sequence

The original time-series data is transformed into a supervised format using a **sliding window approach**, where sequences of past observations are used to predict future values. Specifically:

$$X^i = [x^i, x^{i+1}, ..., x^{i+n-1}], \quad y^i = x^{i+n} \qquad [5]$$

Here, n represents the number of historical observations (timesteps). This transformation enables the model to learn from temporal patterns and is critical for effectively training recurrent networks [6].

D. Feature selection

A univariate input strategy was adopted for all models using the energy consumption value $x_t = E_t$. To ensure clean and consistent input, missing values and zero entries were removed, and interpolation was applied where necessary. The cleaned data was normalized using **Min-Max scaling**, computed as[12]:

$$x^{scaled} = (x - x_{min})/(x_{max} - x_{min}) \qquad [6]$$

This step ensures all features are scaled to the range [0,1], which aids in stabilizing training dynamics and accelerating convergence [6].

E. Hyperparameter selection

Hyperparameters were selected manually based on preliminary experimentation and prior studies. The number of timesteps was varied across three configurations n=2, n=6, and n=12, to evaluate the models' performance over short, medium, and longer-term historical windows. These settings determined how many prior time steps were used as input to forecast the next value. Additional hyperparameters included a batch size of 32 and the use of the Adam optimizer with default settings.

While the original study employed **Mean Absolute Percentage Error (MAPE)** as the loss function, this work adopts **Mean Squared Error (MSE)[7]** during model training.

MSE is particularly advantageous as it places a **greater penalty on larger errors**, which is beneficial for capturing unexpected load spikes or rapid demand fluctuations. Unlike MAPE, which can become unstable when actual values are near zero, MSE remains well-defined across all prediction ranges and offers **smoother gradients for optimization**, leading to more stable and consistent model convergence.

$$MSE = 1/N \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \qquad \forall \ i \in N \qquad [7]$$

Where, y_i is the real value and $\widehat{y_i}$ is the predicted value This choice of loss function is standard in regression-based

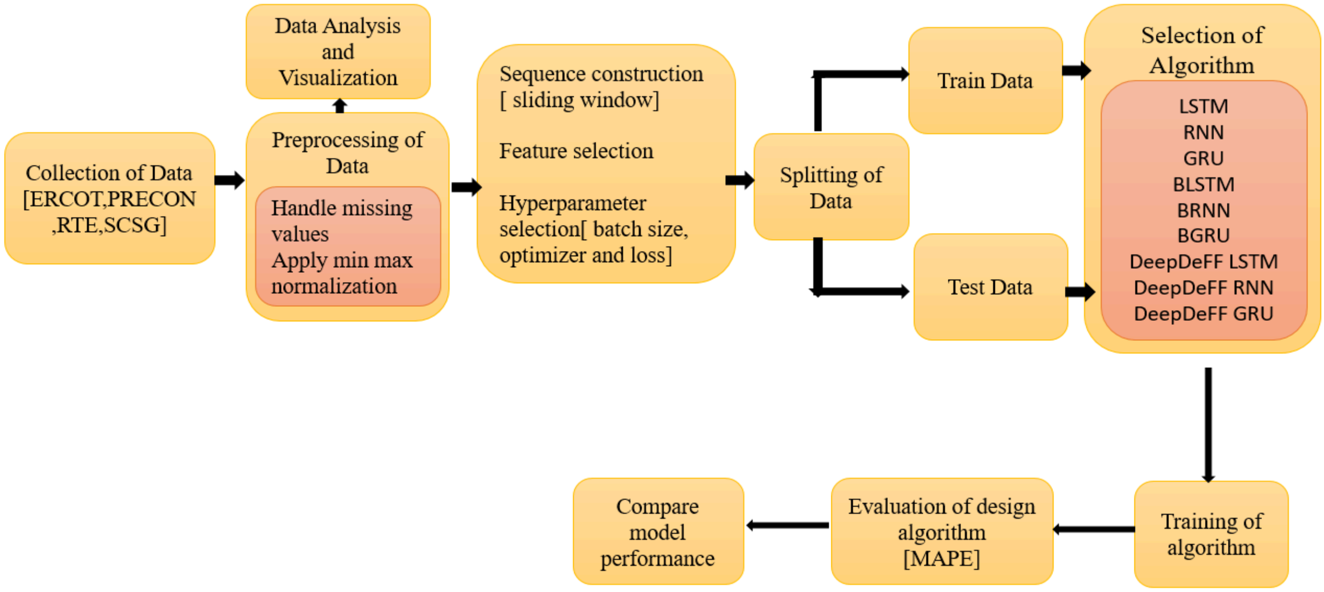
Data Analysis and Visualization
Collection of Data [ERCOT,PRECON ,RTE,SCSG]
Preprocessing of Data
Handle missing values
Apply min max normalization
Sequence construction [sliding window]
Feature selection
Hyperparameter selection[batch size, optimizer and loss]
Splitting of Data
Train Data
Test Data
Selection of Algorithm
LSTM
RNN
GRU
BLSTM
BRNN
BGRU
DeepDeFF LSTM
DeepDeFF RNN
DeepDeFF GRU
Training of algorithm
Evaluation of design algorithm [MAPE]
Compare model performance

Fig 4: proposed system architecture

forecasting tasks and penalizes large deviations between predicted and actual values [6]. No automated hyperparameter optimization was conducted, as the study emphasizes architecture evaluation rather than fine-tuning.

F. Evaluation

In alignment with the original study, this replication also adopts **Mean Absolute Percentage Error (MAPE)** as the sole evaluation metric for forecasting performance. MAPE is widely used in time-series forecasting due to its **scale-independent nature**, ease of interpretation, and direct expression of error as a percentage. It facilitates fair comparison across models and datasets with different scales, making it especially suitable for electricity demand forecasting where values can vary significantly. Additionally, by using the same evaluation criterion as the original paper, this work ensures a consistent and fair comparison of model performance across both studies..

Mean Absolute Percentage Error (MAPE expresses the prediction error as a percentage of the actual value, providing an intuitive measure of relative accuracy.

$$MAPE \;=\; 100/n \sum_{t=1}^{n} \left| \frac{y_i - \widehat{y}_i}{y_i} \right| \qquad [8]$$

All models were trained and tested using an **80/20 train-test split**, and performance was evaluated across different datasets and forecast horizons (e.g., 2, 6, and 12-time steps ahead). These metrics enabled a standardized comparison of model behavior across architectures, timestamp configurations, and data sources.

The complete forecasting pipeline, as outlined in **Fig. 4**, illustrates the end-to-end architecture followed in this study from data collection to model evaluation. This structured approach ensures consistent treatment across datasets and enables a fair comparison of multiple sequential models under varying timestep configurations. With the methodological foundation established and models trained across all datasets, the subsequent section presents detailed results and comparative insights based on the defined evaluation metrics.

IV. The Dataset

To comprehensively evaluate model performance across diverse consumption profiles, this study utilizes four real-world electricity demand datasets: **ERCOT(The Electricity Reliability Council of Texas)**, **PRECON(Pakistan Residential Electricity Consumption)**, and **RTE(Réseau de Transport d'Électricité)**. These datasets were also adopted in the original study by Wahab et al. [6], which serves as the foundation for this replication and extension effort. Each dataset captures different temporal, geographical, and operational characteristics, offering a robust testbed for evaluating sequential deep-learning models.

While the original paper [6] used standardized fixed-length segments of each dataset, this study works with **variable-length data** as available from the respective sources, reflecting realistic constraints in practical deployments. Dataset durations and sampling intervals were carefully retained as-is to preserve natural load patterns and system variability.

A. ERCOT (Electric Reliability Council of Texas) Dataset

The ERCOT dataset[13] consists of **hourly electricity load data** collected from residential customers within the ERCOT power system in Texas. As described in Wahab et al. [1], this dataset was employed due to its **real-world variability, clear periodic patterns, and significant intra-day and seasonal fluctuations**, making it suitable for evaluating short-term load forecasting (STLF) models across different temporal resolutions.

In this study, we adopt a more comprehensive scope by utilizing **the entire year of 2019**, amounting to **8,760 hourly records**. This extended time span provides rich seasonal coverage and increases the model's exposure to diverse consumption patterns, including weekdays, weekends, holidays, and weather-related load variations.

B. PRECON(Pakistan Residential Electricity Consumption)

The PRECON dataset [14] records **minute-level electricity consumption data** from **42 households** in a developing country, representing a wide range of **financial backgrounds, appliance ownership, and daily usage routines**. The original dataset spans from **June 1, 2018 to September 31, 2019**, and was designed to capture nuanced consumption behavior in environments affected by common challenges such as **power outages**. These outages are evident in the raw data through extended intervals of zero consumption.

In alignment with Wahab et al. [6], we utilize the **"E"** variable as the primary input feature, which in this context represents **instantaneous power usage in kilowatts (kW)**. For this study, a **clean, uninterrupted 12-month segment** was extracted, ranging from **June 1, 2018 to May 31, 2019**, resulting in approximately **525,600 data points**. This segment was selected for its completeness and consistent coverage across all minutes of the year.

By preserving the original resolution and variability, the PRECON dataset in this implementation offers a **valuable case study in real-world STLF performance under challenging grid conditions**, including outages and non-uniform load behaviors.

C. Réseau de Transport d'Électricité(RTE) France Dataset

The RTE dataset [15], provided by the French national transmission system operator, contains real-time measurements of **national electricity consumption**. It reflects aggregate demand across residential, commercial, and industrial sectors, offering a stable and smooth load profile. As noted in Wahab et al. [6], the RTE dataset is valuable for benchmarking short-term load forecasting (STLF) models under **macro-level, low-variance conditions**.

In this study, a full year of data from **January 1, 2020 to December 31, 2020** was used, covering the French power grid. The original data—recorded at **15-minute intervals**—was **resampled to an hourly resolution** to maintain consistency with other datasets and reduce computational complexity. After resampling, the dataset contains approximately **8,760 data points**.

V. Results

The evaluation framework developed in this study is the result of an iterative process involving extensive experimentation across three real-world electricity load datasets: **ERCOT**, **PRECON**, and **RTE**. Each experiment was carefully structured to maintain consistency in model architecture, data preprocessing, and evaluation methodology. The forecasting models investigated include six well-established sequential architectures: **RNN**, **GRU**, and **LSTM**, along with their bidirectional variants—**BRNN**, **BGRU**, and **BLSTM**—selected for their effectiveness in time-series modeling and to enable direct comparison with prior work, Wahab et al. [6].

All sequential models share a standardized architectural design,. For consistency and realism, for the process of load forecasting the all **"actual electricity consumption data[E]"** is used, without any synthetic feature augmentation. Prior to training, the data was preprocessed by **removing missing and anomalous values**, followed by **Min-Max normalization** to ensure uniform scaling. Additionally, for the **RTE dataset**, the original 15-minute resolution data was **resampled to hourly intervals** to align with the temporal granularity of the other datasets. Each dataset spans a **complete, uninterrupted year**, thereby preserving key temporal structures such as daily, weekly, and seasonal demand cycles. This uniform data coverage provides a robust foundation for comparative analysis under practical and representative load conditions.

Each dataset was evaluated across **three forecasting horizons—2, 6, and 12 timesteps—corresponding to short-, medium-, and longer-term prediction intervals**. This design, consistent with the original DeepDeFF study, enables a structured analysis of how each model responds to varying temporal dependencies. Shorter horizons assess responsiveness to recent trends, while longer horizons test the model's ability to capture extended sequential patterns. Evaluating across multiple timesteps provides a more **comprehensive and detailed understanding of model behavior**, especially when applied to datasets with different levels of data resolution and variability. The results are organized by dataset and timestep, followed by a consolidated comparison that highlights key performance trends across the entire experimental framework.

A. ERCOT (Electric Reliability Council of Texas) Dataset

The ERCOT dataset, characterized by high daily and seasonal demand variation, presents a realistic test case for short-term load forecasting. The data was chronologically split into a **training set (80%)** and a **test set (20%)**, where the test set was also used as validation data during training.

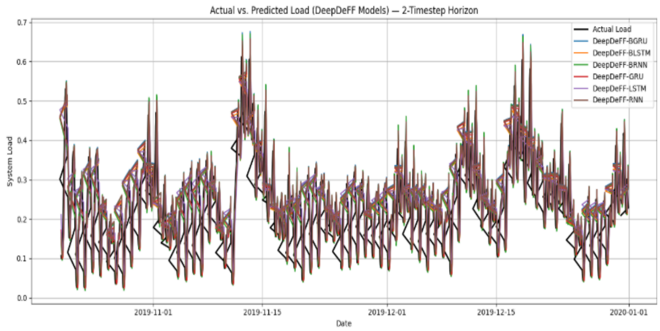

Fig 5: Actual versus predicted system load using DeepDeFF models at a 2-timestep horizon for ERCOT dataset

The Fig 5, illustrates the actual versus predicted system load using DeepDeFF models at a 2-timestamp horizon for the ERCOT dataset. As observed, bidirectional models such as DeepDeFF-BRNN and DeepDeFF-BGRU closely track the real load trends, particularly during peak periods. This supports the strong numerical performance observed in the results, confirming that DeepDeFF models can effectively learn daily and seasonal variations in datasets with moderate variability like ERCOT.

As shown in **Table 1**, model performance on the ERCOT dataset is strongly influenced by the forecasting horizon. While unidirectional models such as **RNN** and **GRU** perform adequately at shorter horizons, their accuracy diminishes with increasing timesteps. **Bidirectional models**, particularly **BRNN** and **BGRU**, demonstrate consistently stronger generalization, with **DeepDeFF-BRNN** achieving the best overall performance—recording the lowest MAPE of **4.52%** at the 12-timestep forecast. In contrast, traditional LSTM and its DeepDeFF variant underperform across all horizons, highlighting their relative limitations in capturing the dataset's temporal variability. The results further confirm the advantage of bidirectional architectures and deep feature fusion in improving forecast accuracy across both short- and long-term horizons

TABLE 1: Results achieved on ERCOT dataset

Method	**Evaluation**		
	2-timestamps	**6- timestamps**	**12-timestamp**
BGRU	MAPE: 6.6%	MAPE:7.21%	MAPE:7.91%
BLSTM	MAPE:10.24%	MAPE:10.28%	MAPE:11.64%
BRNN	MAPE:5.91%	MAPE:**4.62%**	MAPE:7.37%
GRU	MAPE:8.41%	MAPE:8.51%	MAPE:9.52%
LSTM	MAPE:16.19%	MAPE:15.75%	MAPE:10.10%
RNN	MAPE**:4.79%**	MAPE:7.28%	MAPE:8.58%
DeepDeFF LSTM	MAPE: 15.74%	MAPE: 16.82%	MAPE:10.02%
DeepDeFF RNN	MAPE: 6.72%	MAPE:6.11 %	MAPE:8.91 %
DeepDEFF GRU	MAPE: 10.79%	MAPE:7.78%	MAPE: 8.02%
DeepDeFF BLSTM	MAPE:9.71%	MAPE:17.53%	MAPE:11.30%
DeepDeFF BGRU	MAPE:6.24%	MAPE:6.73%	MAPE:8.19%
DeepDeFF BRNN	MAPE:5.18%	MAPE:5.26%	MAPE:**4.52%**

B. PRECON (Pakistan Residential Electricity Consumption) Dataset

The PRECON dataset, collected from 42 households across varying socioeconomic classes, represents a highly diverse and noisy load profile environment. The dataset captures real-world complications such as power outages and irregular usage patterns, making it a valuable test case for evaluating model robustness under uncertain and imbalanced conditions. For this experiment, the data was chronologically split into an 80% training set and a 20% test set, with the test portion also serving as validation during model training.

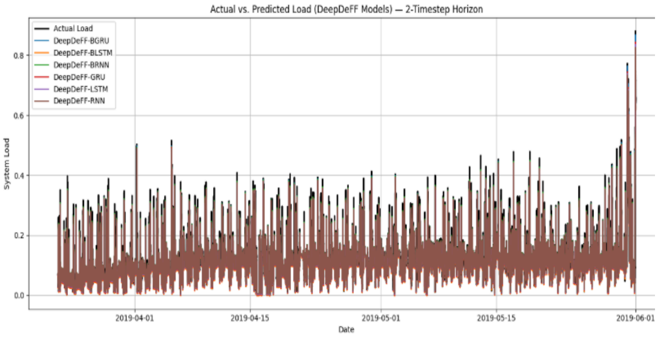

Fig 6: Actual versus predicted system load using DeepDeFF models at a 2-timestep horizon for PRECON dataset

Fig 6 shows the actual vs. predicted system load for the PRECON dataset using various DeepDeFF models over a 2-timestep horizon. All models track the load pattern closely, with bidirectional models like BLSTM and BRNN showing better alignment during sharp load changes. The results highlight the effectiveness of DeepDeFF architectures in capturing short-term variations in the PRECON dataset.

TABLE 2: Results achieved on PRECON dataset

Method	**Evaluation**		
	2-timestamps	**6- timestamps**	**12-timestamp**
BGRU	MAPE:16.62 %	MAPE:22.25%	MAPE:20.0%
BLSTM	MAPE:22.67%	MAPE37.2:%	MAPE:14.06%
BRNN	MAPE36.33:%	MAPE:72.9%	MAPE:20.9%
GRU	MAPE:17.91%	MAPE:21.62%	MAPE:31.72%
LSTM	MAPE:46.56%	MAPE:45.76%	MAPE:66.34%
RNN	MAPE:22.67%	MAPE:**11.07%**	MAPE:18.92%
DeepDeFF LSTM	MAPE:**11.91%**	MAPE: 25.23%	MAPE: 34.43%
DeepDeFF RNN	MAPE:33.18%	MAPE: 28.47%	MAPE: 34.26%
DeepDeFF GRU	MAPE:40.02%	MAPE: 58.19%	MAPE: 23.09%
DeepDeFF BLSTM	MAPE:27.83%	MAPE:25.42%	MAPE:**13.87%**
DeepDEFF BRNN	MAPE:28.93%	MAPE:12.05%	MAPE:14.13%
DeepDeFF BGRU	MAPE:15.74%	MAPE:28.62%	MAPE:29.28%

As shown in **Table 2**, forecasting on the PRECON dataset is notably more challenging due to its higher noise and variability. Performance fluctuates significantly across models and timesteps. Among all models, **DeepDeFF-BLSTM** achieves the best result with a MAPE of **13.87%** at 12 timesteps, followed closely by **DeepDeFF-BRNN**. Traditional **LSTM** and **GRU** models perform poorly, especially at longer horizons. Bidirectional and DeepDeFF variants demonstrate more stable results across horizons, indicating their robustness under complex and irregular load conditions..

C. *RTE(*Réseau de Transport d'Électricité) *Dataset*

The RTE dataset, representing national-level electricity demand at hourly intervals, offers smoother load profiles with minimal short-term fluctuations. This stability supports highly accurate forecasts across all models. Data was chronologically split into 80% training and 20% testing, with the test set also serving as validation during training.

TABLE 3: Results achieved on RTE dataset

Method	Evaluation		
	2-timestamps	**6- timestamps**	**12-timestamp**
BGRU	MAPE: **2.37%**	MAPE:2.74%	MAPE:**2.37%**
BLSTM	MAPE:2.69%	MAPE:3.42%	MAPE:2.94%
BRNN	MAPE:2.44%	MAPE:2.35%	MAPE:**2.37%**
GRU	MAPE:3.80%	MAPE:2.52%	MAPE:2.87%
LSTM	MAPE:6.50%	MAPE:5.80%	MAPE:7.85%
RNN	MAPE:5.97%	MAPE:**2.19%**	MAPE:3.16%
DeepDeff RNN	MAPE: 5.06%	MAPE:5.81%	MAPE:6.54%
DeepDeff LSTM	MAPE:3.20%	MAPE:3.07%	MAPE:3.11%
DeepDeff GRU	MAPE:2.54%	MAPE:2.79%	MAPE:2.88%
DeepDeFF BRNN	MAPE:3.18%	MAPE:4.19%	MAPE:5.03%
DeepDeFF BLSTM	MAPE:3.53%	MAPE:3.27%	MAPE:3.70%
DeepDeFF BGRU	MAPE:3.69%	MAPE:3.19%	MAPE:4.09%

As presented in **Table 3**, forecasting performance on the RTE dataset is consistently strong across all models, reflecting the dataset's smoother and less volatile load profile. Traditional models like **BGRU**, **BRNN**, and **BLSTM** maintain low MAPE values across all horizons, with **BRNN** achieving the best score of **2.35%** at the 6-timestep horizon. Among DeepDeFF variants, **DeepDeFF-GRU** shows the most consistent accuracy, outperforming its baseline counterpart across all timesteps. Overall, both bidirectional and DeepDeFF models demonstrate reliable performance, confirming their suitability for stable, aggregated datasets like RTE.

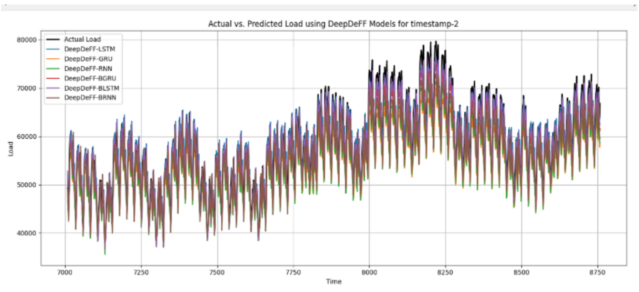

Fig 6: Actual versus predicted system load using DeepDeFF models at a 2-timestep horizon for RTE dataset

The RTE dataset, known for its stable load patterns, enables accurate predictions across all DeepDeFF models. The Fig 6, clearly shows that DeepDeFF-GRU consistently follows the actual load curve with minimal deviation. This visual consistency reinforces the low MAPE values obtained and demonstrates the reliability of DeepDeFF models in low-variance forecasting environments.

VI. Discussion

As seen from the results in Section V, our replicated methodology shows strong and consistent performance, supporting the idea from the original DeepDeFF paper [6] that combining hand-crafted features improves forecasting accuracy. While many earlier works focus on a single dataset, Wahab et al. [6] compared results across five datasets from different countries. In our work, we selected three of those datasets—ERCOT, PRECON, and RTE—and replicated the original framework with a few enhancements, such as using all six sequential models and evaluating across multiple forecasting horizons.

The results also highlight that not all datasets behave the same when it comes to short-term load forecasting (SLF). For instance, individual household data like PRECON shows much higher variability in load consumption compared to large-scale datasets like RTE or ERCOT, which represent state or national-level demand. This high variance makes it harder for forecasting models to learn consistent patterns. One way to reduce this challenge is by using data collected over longer periods—such as a full year—which captures multiple seasons and allows the model to better understand and adapt to the variations in consumption trends over time.

Wahab et al. [6] primarily concentrated on the SGSC and AMPds datasets, where they evaluated performance across multiple forecasting horizons and concluded that the 2-timestep horizon yielded the best results. For the remaining three datasets, they limited their evaluation to only the 2-timestep setting. In contrast, our work extends this analysis by implementing and evaluating all three forecasting horizons (2, 6, and 12 timesteps) across the selected three datasets, thereby offering a more comprehensive performance comparison and enhancing the overall depth of the study.

A. ERCOT dataset

In the ERCOT dataset, model performance varies significantly with the forecasting horizon. At the shortest horizon (2-timestamp), the RNN model delivers the best performance with a MAPE of 4.79%, indicating that traditional unidirectional models can effectively capture short-term dependencies in stable datasets. As the forecasting horizon increases, the need for richer temporal context becomes more evident. At the 6-timestamp level, BRNN achieves the lowest MAPE of 4.62%, demonstrating that bidirectional models offer improved generalization by leveraging both past and future information. Notably, at the longest horizon of 12 timesteps, the DeepDeFF-BRNN model outperforms all others, achieving a MAPE of just 4.52%. This progression highlights an important trend: while simple recurrent models may suffice for near-term forecasting, more complex architectures—such as DeepDeFF combined with bidirectional layers—are better suited for capturing longer-range temporal dependencies. This reinforces the advantage of incorporating deeper and more contextual learning structures when tackling extended forecasting tasks in dynamic systems like ERCOT.

B. RTE dataset

The RTE dataset, characterized by its smooth load profile and low variance, provides a favorable environment for short-term load forecasting models. At a 2-timestep horizon, BGRU demonstrates the best performance with a MAPE of 2.37%, closely followed by other bidirectional models such as BLSTM (2.69%) and BRNN (2.44%), as well as DeepDeFF-GRU (2.54%). This consistency across multiple models suggests that in stable datasets like RTE, several architectures are capable of achieving high accuracy even without extensive temporal context.

At the 6-timestep level, RNN achieves the lowest MAPE of 2.19%, showcasing that even simpler unidirectional models can perform well when the data variance is minimal. However, bidirectional models continue to show strong results—BRNN (2.35%), BGRU (2.74%), and GRU (2.52%) all maintain low error rates. Importantly, DeepDeFF-GRU also delivers a competitive MAPE of 2.79%, reaffirming its robustness in capturing extended temporal patterns.

At the 12-timestep horizon, BGRU and BRNN once again lead with a MAPE of 2.37%, while GRU (2.87%) and DeepDeFF-GRU (2.88%) follow closely. The consistent performance of DeepDeFF-GRU across all three horizons—despite not always being the absolute best—highlights its adaptability and stability in low-variance scenarios. This indicates that while bidirectional models are often optimal for longer-term forecasting, DeepDeFF-GRU maintains competitive accuracy across all horizons, making it a reliable choice for balanced performance in structured datasets like RTE.

C. PRECON dataset

The PRECON dataset presents unique challenges in load forecasting due to its household-level granularity, high variability, and frequent occurrences of 0 kW readings caused by power outages. These zero-load values can significantly distort MAPE calculations due to divide-by-zero issues. To address this, the original study by Wahab et al. [6] proposed a pragmatic approach by replacing 0 kW values with 0.1 kW, minimizing their impact while preserving data integrity. Our study adopts a similar strategy, enabling fair benchmarking across different forecasting horizons.

The PRECON dataset comprises energy consumption data from 42 individual households. While Wahab et al. evaluated the dataset collectively, comparing multiple households together, our study narrows the focus to a single household. This approach enables a more controlled analysis across three forecasting horizons—2, 6, and 12 timesteps—allowing us to closely observe how model performance changes with increasing temporal depth. At the 2-timestep level, DeepDeFF-LSTM achieved the best performance with a MAPE of 11.91%, indicating its effectiveness in capturing immediate load patterns. For the 6-timestep horizon, RNN slightly outperformed other models with a MAPE of 11.07%, suggesting that simpler architectures can still perform well with moderate temporal depth in noisy data. However, as the forecasting window extends to 12 timesteps, DeepDeFF-BLSTM emerged as the most robust model, delivering a MAPE of 13.87%. This trend reaffirms that bidirectional DeepDeFF models are more capable of handling long-term dependencies and irregular consumption behavior.

D. Comparison

In the original study by Wahab et al., smooth and low-variance datasets like ERCOT and RTE showed exceptional forecasting performance using DeepDeFF architectures. At the 2-timestep horizon, DeepDeFF-GRU and DeepDeFF-BGRU achieved MAPE values as low as 0.81% and 0.91%, respectively—demonstrating that even unidirectional DeepDeFF models can perform effectively under stable conditions. In our replication, ERCOT was evaluated over 2, 6, and 12 timesteps, where DeepDeFF-BGRU achieved the best result at 2 timesteps (MAPE: 6.24%), DeepDeFF-RNN at 6 timesteps (MAPE: 6.11%), and DeepDeFF-BRNN at 12 timesteps (MAPE: 4.52%). The higher MAPE values in timestamp 2 compared to the original are attributed to differences in dataset size. For the RTE dataset, our findings align with the original study—DeepDeFF-GRU consistently delivered top performance across all horizons, with MAPE values of 2.54%, 2.79%, and 2.88%, further confirming its robustness in low-noise environments. In contrast, the PRECON dataset presented a more challenging scenario due to high granularity and frequent power outages. Wahab et al. reported MAPE values ranging from 7.67% to 37.61% across various households, with DeepDeFF-BGRU and BRNN performing best. Our study focused specifically on House 42, where DeepDeFF-LSTM achieved a MAPE of 11.91% at 2 timesteps, DeepDeFF-BRNN 12.05% at 6 timesteps, and DeepDeFF-BLSTM 13.87% at 12 timesteps. These results fall within the range reported in the original study and reinforce the DeepDeFF architecture's capability to adapt across varying levels of data complexity.

TABLE 4: Comparative Performance of DeepDeFF Models at 2-Timestamp Horizon Across Datasets

ERCOT	Model	MAPE(%)
original paper	DeepDeFF BGRU	0.91%
Replicated Study	DeepDeFF-BGRU	6.24%
RTE		
original paper	DeepDeFF-GRU	0.81%
Replicated Study	DeepDeFF-GRU	2.54%
PRECON		
Original paper	DeepDeFF-BGRU	21.87%
Replicated Study	DeepDeFF-LSTM	11.91%

Table 4 highlights the best-performing DeepDeFF model for each dataset at the 2-timestep horizon, comparing the original paper's results with those obtained in our replicated study. Differences in MAPE values reflect variations in dataset length, preprocessing, and model configurations.

VII. Future work

This work presents several promising directions for future extension. One potential improvement involves optimizing hyperparameters individually for each dataset and incorporating additional contextual features—such as temperature, humidity, seasonality, and holidays—to strengthen the DeepDeFF model, particularly in cases with unpredictable spikes like those seen in the PRECON dataset. The model's performance is also closely tied to the quality and consistency of the data. In datasets with highly variable or disjoint patterns between training and test sets, such as PRECON or SGSC, forecasting accuracy may decline. Addressing this issue could involve expanding the training period to include an entire year, thereby capturing broader seasonal trends and behavioral variations. Furthermore, experimenting with alternative loss functions may enhance the model's ability to generalize across diverse conditions and improve overall performance.

VIII. Conclusion

This study replicates and extends the DeepDeFF framework by evaluating six sequential models—RNN, LSTM, GRU, and their bidirectional counterparts—across three real-world datasets (ERCOT, PRECON, and RTE) and three forecast horizons (2, 6, and 12 timesteps). The results confirm that bidirectional and DeepDeFF models, particularly DeepDeFF-GRU and DeepDeFF-BRNN, consistently outperform traditional architectures, especially at longer horizons and in stable datasets like RTE. While forecasting remains more challenging in noisy datasets like PRECON, the DeepDeFF variants demonstrate improved generalization and stability. Overall, the study validates the robustness of DeepDeFF under diverse load profiles and forecasting conditions.

IX. References

[1] J. Huan et al., "Short-Term Load Forecasting of Integrated Energy Systems Based on Deep Learning," 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE), Chengdu, China, 2020, pp. 16-20, doi: 10.1109/ACPEE48638.2020.9136566.

[2] M. Songkin et al., "Study of Short-Term Load Forecasting Techniques," 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri Sarawak, Malaysia, 2024, pp. 272-276, doi:10.1109/GECOST60902.2024.10474795.

[3] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu and Y. Zhang, "Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," in IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841-851, Jan. 2019, doi:10.1109/TSG.2017.2753802.

[4] Z. Yang, T. Hu, J. Zhu, W. Shang, Y. Guo and A. Foley, "Hierarchical High-Resolution Load Forecasting for Electric Vehicle Charging: A Deep Learning Approach," in IEEE Journal of Emerging and Selected Topics in Industrial Electronics, vol. 4, no. 1, pp. 118-127, Jan. 2023, doi: 10.1109/JESTIE.2022.3218257.

[5] M. Sayadlou, M. S. Naderi, M. Abedi, S. Esmaeili and M. Amini, "A Comprehensive Deep Learning Method for Short-Term Load Forecasting," 2022 30th International Conference on Electrical Engineering (ICEE), Tehran, Iran, Islamic Republic of, 2022, pp. 1074-1078, doi: 10.1109/ICEE55646.2022.9827325

[6] A. Wahab, M. A. Tahir, N. Iqbal, A. Ul-Hasan, F. Shafait and S. M. Raza Kazmi, "A Novel Technique for Short-Term Load Forecasting Using Sequential Models and Feature Engineering," in IEEE Access, vol. 9, pp. 96221-96232, 2021, doi: 10.1109/ACCESS.2021.3093481.

[7] A. S. Mahajan and A. Shrivastav, "Short Term Load Forecasting based on Regression models," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-8, doi: 10.1109/ICONAT57137.2023.10080359.

[8] R. A. Khan, C. L. Dewangan, S. C. Srivastava and S. Chakrabarti, "Short Term Load Forecasting using SVM Models," 2018 IEEE 8th Power India International Conference (PIICON), Kurukshetra, India, 2018, pp. 1-5, doi: 10.1109/POWERI.2018.8704366.

[9] S. Tsegaye, S. Padmanaban, L. B. Tjernberg and K. A. Fante, "Short-Term Load Forecasting for Electrical Power Distribution Systems Using Enhanced Deep Neural Networks," in IEEE Access, vol. 12, pp. 186856-186871, 2024, doi: 10.1109/ACCESS.2024.3432647.

[10] Y. Feng, G. Xue, H. Li and Y. Zhao, "Short-Term Electricity Load Forecasting Model Based on Bidirectional Long Short-Term Memory Network with Adaptive Boosting," 2024 11th International Forum on Electrical Engineering and Automation (IFEEA), Shenzhen, China, 2024, pp. 433-437, doi: 10.1109/IFEEA64237.2024.10878590.

[11] Y. Zhu, C. Sun, W. Zheng, Y. Gao and X. Zhao, "Short-term Load Forecasting Based on Kmeans and FR-DBN Models," 2020 10th International Conference on Power and Energy Systems (ICPES), Chengdu, China, 2020, pp. 145-150, doi: 10.1109/ICPES51309.2020.9349644.

[12] L. Jia, G. Li, Z. Zhang, Y. Wang, Y. Sun and S. Li, "Deep learning-based short-term load forecasting for power grids," 2024 4th International Conference on Energy, Power and Electrical Engineering (EPEE), Wuhan, China, 2024, pp. 188-191, doi: 10.1109/EPEE63731.2024.10875100.

[13] ERCOT.(2019).GridData.Accessed:Aug.27,2019.[Online]. Available: https://ercot.com/

[14] A. Nadeem and N. Arshad, "PRECON: Pakistan Residential Electricity Consumption Dataset," *Proceedings of the 10th ACM International Conference on Future Energy Systems (e-Energy '19)*, Phoenix, AZ, USA, pp. 52–57, 2019. DOI: 10.1145/3307772.3328317

[15] RTE. (2019). Grid Data. Accessed: Aug. 27, 2019. [Online]. Available: https://data.rte-france.com/