Review

LSTM vs. Transformer Models in Power Forecasting: A Comprehensive Survey

Bijay Bastakoti *, Sachin Parajuli *, Bikesh Regmi

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0624.v1

Keywords: machine learning; deep learning; time series forecasting; energy prediction; LSTM; transformer models; smart grids

 doi:10.20944/preprints202603.0624.v1

Review

LSTM vs. Transformer Models in Power Forecasting: A Comprehensive Survey

Bijay Bastakoti [1,*], Sachin Parajuli [1,*] and Bikesh Regmi [2]

[1] Department of Computer Science, University of Texas at Arlington
[2] Department of Computer Science, Morgan State University
* Correspondence: bxb2270@mavs.uta.edu (B.B.); sxp4677@mavs.uta.edu (S.P.)

Abstract

Accurate forecasting of electricity consumption is essential for optimizing energy resources, load balancing, and grid reliability. As urbanization and the integration of renewable energy accelerate, sophisticated forecasting models become indispensable. Long Short-Term Memory (LSTM) networks have long been relied upon for sequential prediction due to their effective memory architecture. More recently, Transformer models—originally developed for Natural Language Processing—have emerged as powerful alternatives, offering enhanced scalability and superior long-range dependency modeling. This survey provides a detailed comparative analysis of and Transformer-based models for electricity usage forecasting. We evaluated 20 peer-reviewed studies by examining forecasting accuracy, scalability, infrastructure compatibility, and deployment viability. Our review finds that while Transformer models excel at long-range, high-resolution forecasting, LSTMs are valuable for lightweight, real-time applications. We also highlight promising hybrid models that integrate both paradigms. Finally, we discuss the critical impact of machine learning infrastructure and propose future research directions to enhance performance and adaptability.

Keywords: machine learning; deep learning; time series forecasting; energy prediction; LSTM; transformer models; smart grids

---

1. Introduction

1.1. Background

Electricity demand forecasting is fundamental to modern energy systems. As smart grids, electric vehicles, and renewable energy sources become increasingly widespread, accurate forecasting is essential for efficient energy distribution and operational planning. The integration of intermittent renewables further adds complexity and necessitates intelligent models that can learn intricate consumption patterns over time.

1.1.1. Motivation

Traditional forecasting methods, such as ARIMA and linear regression, struggle with nonlinearities and evolving data patterns. LSTM models, with their memory cell architecture, address some of these limitations by capturing short- and medium-term dependencies. However, their sequential nature restricts scalability. Transformer models, with parallel processing and self-attention mechanisms, offer a promising alternative for long-range and high-resolution time-series forecasting Zhou et al. (2021); Vaswani et al. (2017). A systematic evaluation of both is needed to guide model selection in diverse forecasting scenarios.

1.1.2. Objective

The primary objective of this study is to compare LSTM and Transformer models for electricity consumption forecasting, focusing on their ability to capture temporal dependencies and generalize

 doi:10.20944/preprints202603.0624.v1

across various consumption patterns. This comparison analyzes the strengths and weaknesses of the two models with respect to accuracy, computational efficiency, scalability, and explainability. Additionally, the study aims to assess how factors such as data infrastructure, machine learning frameworks, and deployment environments influence the practical performance of these models. Finally, the work seeks to identify existing research gaps and propose directions for future improvements in intelligent electricity demand forecasting systems.

2. Literature Review Methodology

2.1. Paper Selection

We reviewed 20 scholarly papers from databases such as IEEE Xplore, ACM Digital Library, MDPI, Springer, and arXiv. Selection criteria included:

- Model type (LSTM, Transformer, Hybrid)
- Forecasting horizon (short-, medium-, or long-term)
- Dataset characteristics (real-world, regional diversity)
- Evaluation metrics and infrastructure considerations

2.2. Review Criteria

Each paper was selected based on:

- Forecasting performance (e.g., MAE, RMSE)
- Scalability and generalizability to different contexts
- Training and inference efficiency
- Interpretability and usability of results
- Infrastructure and, implementation (tools, frameworks, deployment environments)

3. Overview of Forecasting Models

3.1. Long Short-Term Memory (LSTM)

LSTM networks are a type of Recurrent Neural Network (RNN) engineered to manage the vanishing gradient problem in long sequences. Their core innovation lies in memory cells controlled by input, output, and forget gates. These networks excel at handling sequential data with periodicity or seasonality. The Long Short-Term Memory (LSTM) neural network, as illustrated in Figure 1, is a specialized form of Recurrent Neural Network (RNN) designed to overcome the limitations of traditional RNNs in modeling long-term dependencies. In the context of power consumption forecasting, where historical electricity usage data exhibit temporal dependencies influenced by factors such as time of day, seasonality, and weather conditions, the LSTM architecture is particularly well-suited due to its gating mechanisms.

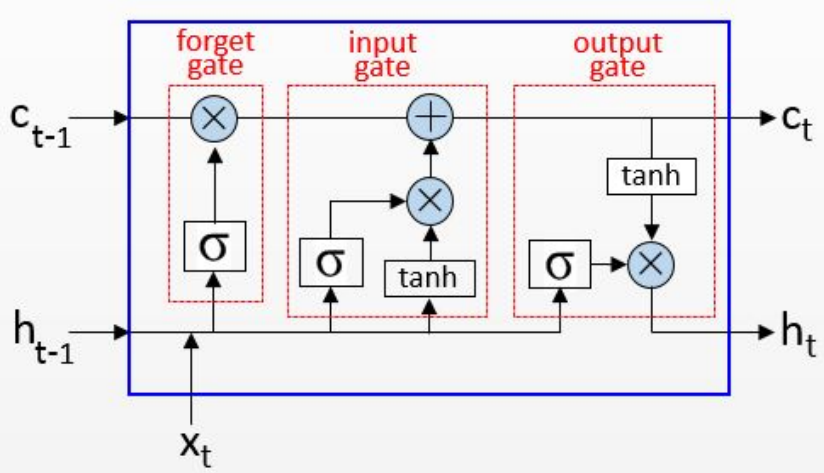

Figure 1. LSTM Architecture.

 doi:10.20944/preprints202603.0624.v1

3.1.1. Working Mechanism of LSTM for Power Forecasting:

The LSTM network processes sequential time-series data, where the input at each time step represents power consumption over a given interval (e.g., hourly or daily). The architecture consists of the following core components that enable effective modeling of such data:

- **Forget Gate**
  The forget gate learns to identify irrelevant information from previous time steps that may not contribute to accurate future power predictions. This is essential in electricity forecasting, where old consumption patterns might become less relevant over time.
- **Input Gate and Candidate Memory Update:**
  The input gate regulates the extent to which new input (e.g., current consumption, temperature, or time features) influences the internal cell state. The candidate memory values propose potential updates to the cell state, enabling the network to learn new consumption patterns, such as demand surges during peak hours or seasonal changes.
- **Cell State Update:**
  By combining the retained long-term memory (via the forget gate) with the relevant new information (via the input gate), the LSTM updates its internal cell state. This ability to carry forward important information across long time horizons helps the model capture periodic behaviors in power usage, such as daily cycles or weekly patterns
- **Output Gate and Hidden State Generation:**
  The output gate determines the contribution of the current cell state to the output prediction. The hidden state generated at each step represents the learned features necessary for forecasting the next step's power consumption.

3.1.2. Advantages and Limitations of LSTM in Power Consumption Forecasting

- **Advantages:**
  The Long Short-Term Memory (LSTM) model demonstrates strong short-term predictive performance, making it highly effective for applications such as power consumption forecasting, where immediate future values are strongly influenced by recent patterns. This strength arises from the LSTM's gated architecture, which allows the network to efficiently capture and process short-term dependencies within sequential data without significant information loss.
  Additionally, LSTM networks are particularly well suited to modeling moderate temporal dependencies, such as daily or weekly cycles often observed in power usage data. The cell state and gate mechanisms enable selective information retention and updating, allowing the model to maintain important context across multiple time steps. This makes LSTM appropriate for scenarios where forecasting depends on both current conditions and recent history.
  Another significant advantage of LSTM models is their compatibility with low-resource environments. Compared with more complex architectures such as Transformer-based models, LSTMs require fewer computational resources and memory, making them deployable on devices with limited processing power or for real-time forecasting applications. This aspect is particularly beneficial for embedded systems or smart meters within distributed energy management infrastructures.
- **Limitations:**
  Despite their strengths, LSTM models exhibit limited scalability due to their inherently sequential computation process. Since each output depends on the previous hidden state, training and inference cannot be fully parallelized across time steps, unlike models based on self-attention. This sequential nature hinders the efficiency of LSTMs, particularly when processing large-scale datasets or high-resolution time-series data.
  Moreover, LSTMs are less efficient at modeling long-range dependencies. Although the gating mechanisms enable better information retention than vanilla RNNs, the ability to capture very long-term patterns—such as seasonal shifts or yearly cycles in power consumption—degrades as

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 March 2026 doi:10.20944/preprints202603.0624.v1

sequence length increases. This limitation becomes critical in long-term load forecasting scenarios where distant historical patterns significantly influence future demand.

Another drawback is the relatively slow training process on large datasets. Due to the step-by-step nature of backpropagation through time (BPTT) and the presence of multiple gate operations at each time step, LSTM networks require substantial training time as the dataset size or sequence length grows. This makes them less suitable for applications that require rapid model updates or large-scale forecasting tasks involving high-dimensional time-series inputs.

3.2. Transformer Models

Transformers leverage multi-head self-attention to compute relationships between all input elements simultaneously. This architecture eliminates recurrence, allowing for parallel computation and scalable learning. Models such as Informer, Autoformer, and Transformer-XL have adapted this architecture for time-series forecasting.

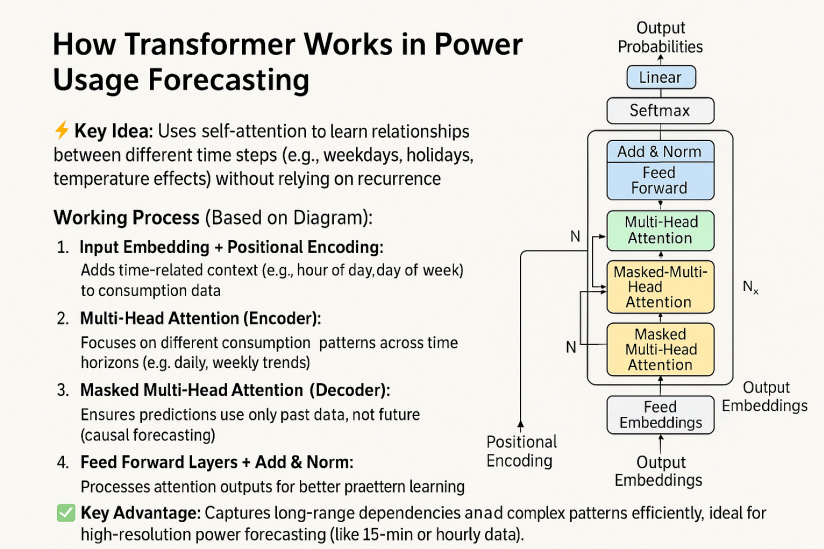

Figure 2. Transformer Architecture.

3.2.1. Advantages and Limitations of Transformer in Power Consumption Forecasting

Advantage:

- Transformer-based models offer **high parallelism and computational efficiency**, making them exceptionally suitable for handling large-scale time-series forecasting tasks, including power consumption prediction. Unlike traditional Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models, which process sequences sequentially, the Transformer architecture employs self-attention mechanisms that allow simultaneous computation across all time steps. This characteristic significantly reduces training time and enhances scalability, particularly for long sequences or large datasets.
- Another key advantage of Transformers is their **superior performance in long-term and multi-variate forecasting scenarios**. The self-attention mechanism enables the model to capture global dependencies across input sequences without the constraints imposed by temporal distance. This allows Transformers to model complex relationships between multiple input features (e.g., power consumption, temperature, humidity, calendar effects) and to maintain high forecasting accuracy over extended prediction horizons, outperforming traditional sequence models in such contexts.
- Furthermore, Transformers provides a **flexible architecture that supports extensive customization**. Variants such as Informer, Autoformer, and Transformer-LSTM hybrids have been developed to address specific challenges in time-series forecasting, including long sequence modeling, seasonal decomposition, and memory efficiency. This adaptability allows researchers and prac-

 doi:10.20944/preprints202603.0624.v1

titioners to tailor Transformer-based models according to the specific needs of their forecasting tasks, including energy load prediction at various temporal resolutions.

LImitation:

- Despite these strengths, Transformer models are **resource-intensive**, requiring significantly higher computational power and memory compared to RNN-based models. The multi-head self-attention mechanism, along with the stacking of encoder-decoder layers, increases both the parameter count and processing overhead. This makes deployment challenging in low-resource environments, such as edge devices or embedded systems, where computational efficiency is critical.
- In addition, Transformer architectures often **require extensive data preprocessing and feature engineering** to achieve optimal performance in power consumption forecasting tasks. Since the model does not inherently account for temporal ordering (unlike RNNs), positional encoding and careful scaling of input variables become necessary to maintain the sequence structure. Handling missing values, normalizing inputs, and selecting appropriate time features (e.g., hour of day, day of week, holidays) are critical steps that can significantly influence forecasting accuracy.
- Another notable limitation is the **increased complexity in model interpretability**. Due to the dense self-attention layers and the high-dimensional representation space, explaining the decision-making process of a Transformer model is more challenging than that of simpler architectures such as LSTM or linear regression models. While attention weights provide some insight into feature importance, they do not always translate directly into clear interpretability for non-expert stakeholders in the energy sector.

3.3. Hybrid Models

Hybrid approaches combine LSTM and Transformer architectures to leverage complementary strengths. CNN-LSTM structures use CNNs for spatial pattern extraction followed by LSTMs for sequential modeling Saoud et al. (2022); Jin et al. (2020). Informer-LSTM integrates Transformer attention with LSTM gates.

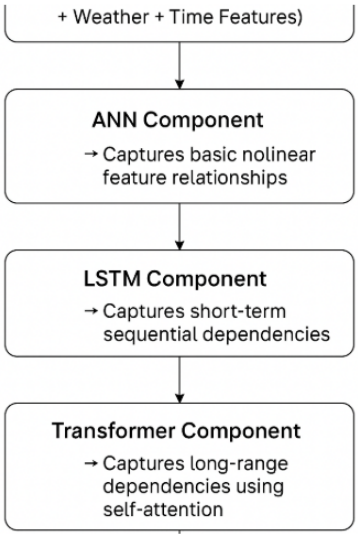

Figure 3. Hybrid Model Architecture.

3.3.1. Working Process

- **Step1: Input Dats Preparation**
  Historical electricity consumption, weather data (temperature, humidity, wind speed), calendar information (day of week, holidays), and other relevant features are collected and preprocessed. Data normalization and feature engineering are applied to prepare the data for model input.

 doi:10.20944/preprints202603.0624.v1

- **Step 2: ANN Component**
  An Artificial Neural Network (ANN) is used to initially process the input features. The ANN captures basic, nonlinear interactions between variables. It transforms the raw input into a more abstract representation for the next layers.
- **Step 3: LSTM Component**
  The ANN output is fed into a Long Short-Term Memory (LSTM) network. LSTM excels at modeling short-term temporal dependencies and sequential patterns in time series data, such as recent usage trends or immediate past effects.
- **Step 4: Transformer Component**
  The LSTM output is passed to a Transformer encoder. The Transformer utilizes a self-attention mechanism to capture long-range dependencies across the entire input sequence. This enables the model to learn from distant historical periods, seasonal trends, or rare events that affect future consumption.
- **Step 5: Output Layer**
  The Transformer output is passed through fully connected layers to produce the final prediction: the forecast electricity consumption or market price for the specified future time horizon.

3.3.2. Advantage of Hybrid Model

- **High Accuracy:** The combination of ANN, LSTM, and Transformer layers allows the model to capture both local and global patterns effectively.
- **Long-Term Dependency Modeling:** The Transformer layer significantly improves the ability to model long-range dependencies compared to LSTM alone.
- **Scalability:** The model can handle large datasets and long input sequences efficiently.
- **Real-world Performance:** Experiments demonstrate that the hybrid model consistently outperforms traditional single-model approaches in terms of forecasting accuracy, robustness, and adaptability to volatile electricity consumption patterns

4. Comparative Analysis

4.1. Forecasting Accuracy

LSTM models are consistent for 24–48-hour forecasts, as demonstrated in Choi et al. (2020), while Transformers tend to outperform in longer-horizon forecasting scenarios, particularly beyond 48 hours Wu et al. (2021). Hybrid models such as LSTM-Informer perform reliably across short-, medium-, and long-term horizons, showing strong versatility across diverse datasets Wang et al. (2023).

Model Type	MAE Range	RMSE Range	Best Use Case
LSTM	0.08–0.15	0.10–0.25	Short-term forecasting
Transformer	0.05–0.12	0.08–0.20	Long-term forecasting
LSTM-Informer	0.04–0.10	0.07–0.18	Hybrid use cases

Figure 4. Forecasting Accuracy.

4.2. Scalability

Scalability is a critical factor in selecting forecasting models for operational energy systems, especially as the volume, velocity, and dimensionality of data continue to increase in smart grids and distributed energy resource environments.

- **LSTM Scalability:** LSTM networks, while effective at modeling sequential dependencies, suffer from limited scalability due to their inherent sequential computation structure. As highlighted

 doi:10.20944/preprints202603.0624.v1

in Choi et al. (2020) and Jin et al. (2020), training LSTM models becomes increasingly time-consuming with longer input sequences, owing to the need for sequential backpropagation through time (BPTT). Their reliance on hidden state propagation across time steps prevents full parallelization, making large-scale training slower and more resource intensive. Furthermore, Zhang et al. (1998) and Jang et al. (2024) observe that LSTM models require substantial fine-tuning of hyperparameters such as sequence length and learning rate to maintain stability on large datasets. In deployment settings, LSTM models are advantageous for small- to medium-scale datasets and for short-term forecasting tasks (e.g., 24–48 hours) where computational efficiency and reduced hardware requirements are priorities.

- **Transformer Scalability:** In contrast, Transformer-based architectures exhibit superior scalability, particularly when trained and deployed on distributed GPU environments. Studies Zhou et al. (2021); Oliveira and Oliveira (2023), andWu et al. (2021) demonstrate that Transformers can handle significantly longer input sequences without performance degradation, thanks to their self-attention mechanisms which allow full parallelization across time steps. For instance, the Informer model proposed in Zhou et al. (2021) introduced a ProbSparse self-attention mechanism that reduces computational complexity from $O(L^2)$ to $O(L \log L)$, where L is the sequence length. This innovation significantly improves training speed and enables efficient modeling of extremely long-term power usage patterns. Similarly,Wu et al. (2021) showed that Autoformer improves scalability by incorporating decomposition blocks that separately model trend and seasonal components. However, it is worth noting that Transformer models still exhibit quadratic memory growth with respect to input length during self-attention calculations Vaswani et al. (2017), which can pose challenges for extremely high-frequency or multi-modal datasets unless memory-efficient variants (e.g., Linformer, Reformer) are employed.
- **Hybrid Model Scalability:** Hybrid architectures combining LSTM and Transformer components, such as the Transformer-LSTM proposed by Qiao et al. (2023) and the CNN-LSTM-Transformer studied by Ahmadipour et al. (2024), aim to balance scalability and sequential modeling. These hybrids often employ convolutional or encoder layers to compress the input space prior to sequential modeling, thereby reducing computational burden. Nevertheless, as highlighted by Saoud et al. (2022), hybrid models typically involve greater architectural complexity, which can increase deployment and maintenance overhead.
- **Comparative Study:** LSTM models scale adequately for low- to moderate-sized datasets and short forecasting horizons but face bottlenecks for long sequences. Transformer models scale excellently across data volume and sequence length, particularly when leveraging distributed GPU clusters and memory-optimized variants. Hybrid models offer a flexible compromise but introduce architectural complexity that must be managed carefully during deployment. As emphasized in Pokhrel et al. (2022); Qureshi et al. (2024), the choice between these architectures ultimately depends on the interplay among forecast-horizon requirements, available infrastructure, and acceptable trade-offs between model accuracy and training efficiency.

4.3. Infrastructure Needs

4.3.1. LSTM Infrastructure Needs

LSTM-based models are generally lightweight and suitable for deployment in resource-constrained environments. As demonstrated by Choi et al. (2020) and Jin et al. (2020). LSTM models can be efficiently trained and deployed on commodity CPUs without requiring specialized hardware accelerators. Their memory footprint is relatively small, owing to a limited number of parameters compared to Transformer-based architectures.

During the training phase, LSTM models can be trained on standard workstation environments with moderate memory (8–32 GB RAM) and without GPU acceleration, although GPU acceleration can accelerate convergence, especially for larger datasets.

At the inference phase, LSTMs are ideal for edge deployments, such as smart meters, IoT devices, or microgrid controllers. Their low computational requirements enable them to operate in low-latency,

 doi:10.20944/preprints202603.0624.v1

real-time settings, making them well suited for distributed energy management systems in which rapid decision-making is crucial.

Examples:

Liu et al. (2017) successfully deployed LSTM-based models in smart city scenarios using only ARM-based microcontrollers.

Qiao et al. (2023) highlight that LSTM-based hybrid models can even be optimized further for embedded deployment by pruning redundant layers.

4.3.2. Transformer Infrastructure Needs

Transformer models require substantially greater infrastructure support during both training and inference. Due to the self-attention mechanism, standard Transformer architectures exhibit quadratic memory growth with respect to input sequence length Vaswani et al. (2017), making training computationally expensive.

During the training phase, as shown by Zhou et al. (2021) and Wu et al. (2021), Transformer-based models such as Informer and Autoformer typically require multi-GPU setups (e.g., NVIDIA V100 or A100 clusters) and high-bandwidth storage systems capable of handling large batches and long sequences. Training on CPU-only systems is often impractical due to prohibitive training times.

Modern frameworks such as Hugging Face Transformers, PyTorch Lightning, and Horovod are commonly used to facilitate distributed training across nodes on cloud platforms, including AWS EC2, Google Cloud TPU pods, and Azure Machine Learning clusters.

During inference, while Transformers offer fast inference on GPUs, their memory requirements remain high. Lightweight variants like Linformer and Informer partially mitigate this by reducing attention complexity, but the Transformer family remains challenging to deploy on resource-limited edge devices without significant model compression.

Examples:

Oliveira and Oliveira (2023) observed that Transformer models scaled efficiently when deployed on cloud-based Kubernetes clusters with autoscaling capabilities. Zhong (2023) demonstrated the need for memory-optimized attention mechanisms when forecasting multi-region electricity consumption to avoid GPU memory bottlenecks.

4.3.3. Hybrid Model Infrastructure Needs

Hybrid architectures combining CNNs, LSTMs, and Transformers Saoud et al. (2022); Ahmadipour et al. (2024) ,and Saeed et al. (2025) present a middle ground in terms of infrastructure demands. While training these models often requires GPUs due to their composite nature, careful architectural design (e.g., compressing inputs via CNNs) can reduce the computational burden. Inference efficiency in hybrid models varies widely and depends on the balance between LSTM and Transformer components. Models that favor LSTM dominance tend to be more edge-compatible, whereas Transformer-heavy hybrids are more suited for cloud deployment

Model Type	Training Hardware	Inference Deployment	Scalability
LSTM	CPU/GPU optional	Edge devices, CPUs	Good for small to medium datasets
Transformer	Multi-GPU, Cloud	Cloud, High-memory servers	Excellent for large datasets and long sequences
Hybrid	GPU recommended	Edge/Cloud depending on design	Flexible but variable

Figure 5. Comparative Summary on Scalability.

 doi:10.20944/preprints202603.0624.v1

4.4. Explainability

Explainability is a crucial consideration when deploying machine learning models for energy forecasting, particularly in critical infrastructure settings where transparency, regulatory compliance, and stakeholder trust are essential. The interpretability of LSTM and Transformer-based models varies substantially due to differences in their internal architectures and computation mechanisms.

4.4.1. LSTM Model Explainability

LSTM models offer **moderate to high levels of interpretability** through the visualization of internal states, namely the memory cell and the three gating mechanisms—input, output, and forget gates.

As shown in Choi et al. (2020) and Jin et al. (2020), tracking LSTM gate activations over time can help infer which input features or time steps are important for forecasting outcomes. For instance, the forget gate's behavior can indicate whether the model is discarding outdated consumption patterns, while the input gate activations show how new information (e.g., sudden temperature changes or holiday effects) is integrated into the memory state. In practical deployments, tools such as **Hidden State Visualizers** and **Saliency Maps for RNNs** have been used to trace information flow through LSTM networks. Furthermore, methods such as **Integrated Gradients** can attribute predictions to specific input features, providing a form of feature-importance analysis.

4.4.2. Transfromer Model Explainability

Transformer models pose greater interpretability challenges, primarily due to their highly parallelized, non-sequential self-attention mechanisms. Unlike LSTM's step-by-step memory updates, Transformers distribute importance weights across all time steps through multi-head attention, making causal analysis less intuitive. Nonetheless, several strategies have been developed to interpret Transformer models in time-series forecasting:

- **Attention Maps:** As first introduced by Vaswani et al. (2017), attention weights between input tokens can be visualized to understand which historical time steps are influencing current predictions. In the energy forecasting context, Oliveira and Oliveira (2023) demonstrated that analyzing attention heads can reveal which days or hours are critical for predicting future consumption.
- **SHAP (SHapley Additive explanations):**SHAP values can be computed for each input feature to assess its contribution to the model's prediction. In time-series applications, temporal SHAP variants allow identification of the most influential time slices or features.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME generates local approximations of the model's behavior near a given prediction, which can be particularly useful when deploying Transformer models in energy systems where individual predictions must be justified Wu et al. (2021).

However, attention-based explanations are not always strictly causal Jang et al. (2024). As noted in recent literature, a high attention score between two time steps does not guarantee causal influence, which limits interpretability relative to classical statistical models or even simpler RNNs.

Aspect	LSTM	Transformer
Interpretability	Moderate to High	Low to Moderate
Internal Mechanism	Gates and Cell States	Multi-head Attention
Methods Available	Hidden State Tracing, Saliency Maps, Integrated Gradients	Attention Maps, SHAP, LIME
Trust for Stakeholders	Easier to justify (cell behavior)	Needs specialized tools for transparency

Figure 6. Comparative Summary on Explainability.

 doi:10.20944/preprints202603.0624.v1

Overall, LSTM models are inherently more interpretable due to their sequential structure and explicit memory representations. Transformer models, while offering superior forecasting accuracy for long horizons, require additional explainability layers to be deployed confidently in operational energy environments. Techniques like attention analysis, SHAP, and LIME are crucial for bridging this gap but still face limitations in causal clarity.// As energy systems move towards greater autonomy and AI-driven decision-making, improving the interpretability of Transformer models remains a significant research priority Qureshi et al. (2024).

4.5. Deployment Consideration

Deploying forecasting models in production environments—whether at the grid scale, in smart homes, or in microgrid controllers—requires careful evaluation of multiple operational factors beyond predictive accuracy. Model choice directly affects inference speed, resource consumption, deployment costs, ease of system integration, and ongoing maintenance requirements.

4.5.1. Inference Speed:

LSTM models offer moderate inference speed, which is generally sufficient for short- to medium-horizon forecasting tasks. Because of their sequential nature, LSTMs process input step by step, resulting in slightly higher latency than fully parallelizable architectures. However, for time-critical applications with small input sequences, LSTM inference latency remains acceptable, even on CPU-based systems Choi et al. (2020).
Transformer models enable fast inference, particularly when deployed on GPUs or TPU accelerators. Their parallel processing of entire sequences enables them to generate predictions rapidly, making them well-suited for applications requiring real-time long-horizon forecasts Zhou et al. (2021); Wu et al. (2021). However, attention calculation over long sequences can still lead to noticeable latency if not optimized.

4.5.2. Hardware and Memory Constraints:

LSTM models are well-suited to deployment on **resource-constrained hardware** such as edge devices, smart meters, or embedded controllers. Their relatively small model size, minimal memory requirements, and lack of specialized accelerators enable cost-effective, decentralized forecasting systems Liu et al. (2017); Dong et al. (2025). In contrast, Transformer models are **memory-intensive**, particularly when dealing with long input sequences. Even at inference time, Transformer-based models require substantial memory bandwidth to perform attention operations. As noted in Oliveira and Oliveira (2023), high-end cloud infrastructure—equipped with GPUs or TPUs—is often necessary to deploy Transformers for real-world energy forecasting at scale

4.5.3. Deployment Cost:

Deployment costs vary significantly between LSTM and Transformer models.

4.6. Lists

- **LSTM deployments** are inexpensive, both in terms of hardware requirements and maintenance. They can often be integrated into existing infrastructure with minimal upgrades, favoring applications in microgrids or rural smart energy systems.
- **Transformer deployments**incur higher upfront and operational costs due to their reliance on specialized cloud instances, GPU clusters, and greater storage needs. However, these costs can be offset by superior forecasting accuracy and scalability in large-scale grid management systems Zhong (2023)

4.6.1. Ease of Integration (Edge vs Cloud):

LSTM models are highly flexible for edge deployments. They can be embedded into ARM-based microcontrollers, Raspberry Pi systems, or industrial gateways, enabling localized forecasting without

 doi:10.20944/preprints202603.0624.v1

reliance on constant network connectivity.

Transformer models, while powerful, are less suited for edge integration without significant optimization (e.g., model pruning, knowledge distillation). Most Transformer deployments are cloud-centric, leveraging platforms like AWS SageMaker, Azure ML, or Google Vertex AI for managed services and autoscaling capabilities Jang et al. (2024) Hybrid architectures offer intermediate flexibility, depending on the proportion of LSTM versus Transformer components.

4.6.2. Monitoring and Maintenance:

Both LSTM and Transformer models benefit from modern MLOps pipelines for:

- Model monitoring (latency, drift detection, retraining triggers)
- Continuous integration/deployment (CI/CD) using tools like MLflow, DVC, and Kubeflow
- Model explainability audits using SHAP, LIME, and attention visualization

However, maintaining Transformer models in production generally involves **higher operational complexity** due to larger model sizes, more frequent retraining needs, and intricate hyperparameter tuning requirements.

Deployment Factor	LSTM	Transformer	Hybrid Models
Inference Speed	Moderate (good for small inputs)	High (with GPU acceleration)	Variable (depends on architecture)
Hardware Requirements	Low (CPU or edge device)	High (GPU/TPU, high memory)	Medium to High
Deployment Cost	Low	High	Medium to High
Integration Flexibility	High (Edge-friendly)	Medium (Cloud-centric)	Variable
Maintenance Complexity	Low to Moderate	High	High (multi-component systems)

Figure 7. Comparative Summary Table.

4.6.3. Final Remarks:

Deployment feasibility plays a critical role in model selection for real-world forecasting applications.

- For low-latency, cost-sensitive, and decentralized systems, LSTM remains highly effective.
- For centralized, large-scale grid management or multi-region forecasting, Transformers offer scalability and predictive performance but require substantial infrastructure investment.
- Hybrid models offer a promising compromise but must be evaluated carefully with respect to operational complexity.

Thus, matching the forecasting horizon, data frequency, and system architecture with the appropriate model deployment strategy is essential for practical success

5. Infrastructure Impact

The performance and deployment efficiency of LSTM and Transformer models for power forecasting are highly dependent on the supporting infrastructure, ranging from data storage and processing pipelines to model training environments and deployment platforms. This section examines how infrastructure decisions influence model effectiveness, scalability, and real-world feasibility.

5.1. Data Storage and Pipelines

Power forecasting models, particularly those using high-frequency data (e.g., 10-minute or 1-minute intervals), benefit from high temporal resolution but impose significant strain on storage and data pipelines. Proper infrastructure is required to ensure timely ingestion, transformation, and availability of data for training and inference.

 doi:10.20944/preprints202603.0624.v1

- **Storage Requirements:** Transformer models, due to their self-attention mechanisms, scale quadratically with input sequence length, which makes memory optimization critical. Systems such as HDFS and object storage (S3, Azure Blob) are often paired with scalable caching and sharding strategies.
- **ETL & Stream Processing:** Frameworks such as Apache Kafka, Flink, and Spark Streaming support real-time data ingestion from smart meters, sensors, and weather APIs. These systems can be configured to support event-driven or batch workflows, which are essential for updating forecasts dynamically.
- **Cloud-Native Workflows:** Tools like AWS Glue, Azure Data Factory, and Google Cloud Dataflow offer low-code orchestration and metadata-driven workflows to automate ETL pipelines. These become vital for Transformer models trained on large datasets from multiple sources Zhong (2023); Aslam et al. (2021)

5.2. Machine Learning Frameworks and Toolchains

The choice of ML frameworks impacts not just ease of development but also training time, interpretability, and deployment portability.

5.2.1. LSTM Models:

- Supported by **TensorFlow, PyTorch, Keras**
- Rapid prototyping and wide compatibility with existing time series toolkits (e.g., GluonTS).
- Easier to optimize on CPU-based systems.
- Smaller memory footprint and fewer hyperparameters to tune.

5.2.2. Transformer Models:

- Supported by **Hugging Face Transformers, Informer, Autoformer, DeepAR, PyTorch Forecasting**
- Require positional encodings, masking layers, and often more epochs to converge
- Benefit significantly from GPU/TPU acceleration
- Integrate well with modern explainability frameworks (e.g., SHAP, LIME for attention maps)

5.3. Edge vs. Cloud Deployment

LSTM and Transformer models differ significantly in their suitability for edge and cloud deployment, depending on hardware constraints, latency requirements, and compute capacity.

- **Edge Devices:** LSTM models are highly suitable for edge deployment due to their lightweight architecture and low memory footprint. They can run efficiently on resource-constrained hardware such as ARM-based microcontrollers or NVIDIA Jetson boards for real-time energy predictions Liu et al. (2017); Shu et al. (2021). This makes them ideal for localized, low-latency applications in smart homes or microgrids that require short-sequence forecasts. In contrast, Transformer models are less suitable for edge use without substantial model compression or optimization, owing to their high computational and memory demands Vaswani et al. (2017); Oliveira and Oliveira (2023).
- **Cloud Platforms:** Transformer models excel in cloud environments where high-performance computing resources—such as GPUs or TPUs—are available. These models benefit from distributed training frameworks such as Horovod or Ray and are often deployed on Kubernetes clusters or MLflow-serving environments Zhou et al. (2021); Jang et al. (2024). Such setups enable scalable, long-horizon, or multi-site energy forecasting. LSTM models, while functional in the cloud, typically offer only moderate scalability benefits in comparison.

5.4. Training Efficiency and Resource Utilization

Model architecture and infrastructure jointly determine training efficiency and scalability:

 doi:10.20944/preprints202603.0624.v1

- **LSTM Training:** Lower GPU usage, relatively fast convergence on small-to-medium datasets. Suitable for small enterprises or pilot studies.
- **Transformer Training:** High memory bandwidth demand and longer training cycles, especially with long input sequences. Efficient training requires optimizers like **AdamW**, mixed precision, and learning rate schedulers.

Some recent work also integrates **Informer-like distillation**, in which attention scores are pruned layer by layer to improve efficiency for long sequences without significant accuracy loss Zhou et al. (2021).

5.5. Model Interpretability and Monitoring

Infrastructure also supports tools for model explainability and lifecycle management:

- **LSTM Models:** Easily interpretable through gate activations and visualizations of cell memory states.
- **Transformer Models:** Attention heatmaps allow insight into which time steps (or sensors) the model is focusing on—crucial for trust in critical infrastructure like power grids.

Monitoring tools such as **Prometheus, TensorBoard,** and **Grafana** can be integrated into MLOps pipelines to track drift, latency, and retraining triggers.

5.6. MLOps Integration and Automation

Efficient power forecasting in real-world settings requires continuous integration and automated deployment cycles:

- **CI/CD Pipelines:** GitHub Actions, Jenkins, and GitLab are commonly used to automate testing, retraining, and deployment of updated models.
- **Model Versioning:** Tools like DVC (Data Version Control) and MLflow help track model performance over time and enable rollback or ensemble management.
- **Infrastructure-as-Code (IaC):** Terraform, CloudFormation, and Pulumi automate cloud resource provisioning, which is particularly useful for scalable Transformer model training.

Figure 8 illustrates a typical pipeline used in power forecasting systems. The process begins with Ingestion, in which raw data, such as electricity usage, weather conditions, and time-based features, are collected from multiple sources. This is followed by a Preprocessing stage, where the collected data undergoes cleaning, normalization, and transformation to prepare it for model input. The final step in the illustrated pipeline is Model Training, in which machine learning or deep learning models (e.g., LSTM, Transformer, or hybrid architectures) are trained on the preprocessed data. This modular pipeline serves as the foundational workflow for developing accurate and robust power consumption forecasting systems and is typically followed by model evaluation and deployment to cloud or edge environments.

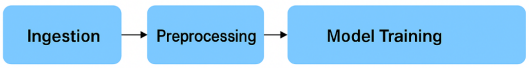

Figure 8. Power Forecasting Pipeline Diagram showing ingestion→ preprocessing → model Training → cloud/edge deployement.

6. Research Gaps and Future Directions

Despite significant advancements, several open challenges and future research opportunities remain in the domain of deep learning-based power consumption forecasting:

6.1. Diverse Dataset Evaluation

Current studies largely rely on datasets from limited regions or sensor infrastructures. There is a lack of cross-domain validation across industries, climates, urban and rural grids, and heterogeneous

 doi:10.20944/preprints202603.0624.v1

sensor platforms. Broader benchmarking across diverse, large-scale, and multimodal datasets is essential for evaluating model generalizability and robustness.

6.2. Explainability Tools

While attention visualization offers partial insights, Transformer models require more intuitive, real-time explanation tools that can be operationalized in critical energy systems. Developing causality-aware attention mechanisms and interpretable sequence modeling techniques remains an important area for future research.

6.3. Resource-Efficient Transformers

The quadratic complexity of self-attention mechanisms imposes challenges for large-scale or real-time forecasting. Lightweight Transformer variants, such as Linformer, Performer, and Informer-based architectures, show promise, but further optimization for resource-constrained environments is necessary to enable widespread deployment.

6.4. Adaptive Forecasting

Energy consumption patterns are dynamic, influenced by behavioral, climatic, and policy changes. Future models should incorporate online learning, concept drift adaptation, and anomaly detection capabilities to maintain forecasting accuracy in evolving environments.

6.5. Multi-Modal Inputs

Incorporating auxiliary features such as temperature, humidity, occupancy rates, grid prices, and policy signals can significantly enhance model performance. Future research should explore robust architectures for integrating structured and unstructured multimodal data streams in power forecasting.

7. Conclusion

This survey presents a comprehensive comparative analysis of LSTM and Transformer models for advanced power consumption forecasting. Through a review of 20 recent studies, we highlight the respective strengths and trade-offs between architectures:

- **LSTM models:**remain effective for short- to medium-term forecasting tasks, particularly in low-resource and real-time operational settings. Their lower computational demands and interpretability make them suitable for edge deployments and decentralized energy systems.
- **Transformer models:**provide state-of-the-art forecasting performance, excelling in long-horizon, high-resolution, and multivariate time series prediction. However, their deployment often necessitates greater infrastructure support and advanced explainability mechanisms.
- **Hybrid models** that integrate LSTM and Transformer components demonstrate promising performance across diverse forecasting horizons, offering potential pathways to balanced accuracy, scalability, and deployment flexibility.

As energy systems continue to evolve toward greater complexity, decentralization, and integration with renewable energy sources, forecasting solutions must not only achieve superior predictive performance but also align with infrastructure constraints, data availability, and operational interpretability requirements. Future research directions including adaptive forecasting, lightweight model architectures, and real-time explainability tools will be critical to advancing the practical deployment of deep learning models in sustainable and resilient energy management systems.

References

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, 2021, [arXiv:cs.LG/2012.07436].

 doi:10.20944/preprints202603.0624.v1

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.

Saoud, L.S.; AlMarzouqi, H.; Hussein, R. Cascaded Deep Hybrid Models for Multistep Household Energy Consumption Forecasting, 2022, [arXiv:cs.LG/2207.02589].

Jin, Y.; Guo, H.; Wang, J.; Song, A. A Hybrid System Based on LSTM for Short-Term Power Load Forecasting. *Energies* **2020**, *13*. https://doi.org/10.3390/en13236241.

Choi, E.; Cho, S.; Kim, D.K. Power Demand Forecasting using Long Short-Term Memory (LSTM) Deep-Learning Model for Monitoring Energy Sustainability. *Sustainability* **2020**, *12*. https://doi.org/10.3390/su12031109.

Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems; Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; Vaughan, J.W., Eds. Curran Associates, Inc., 2021, Vol. 34, pp. 22419–22430.

Wang, K.; Zhang, J.; Li, X.; Zhang, Y. Long-Term Power Load Forecasting Using LSTM-Informer with Ensemble Learning. *Electronics* **2023**, *12*. https://doi.org/10.3390/electronics12102175.

Zhang, G.; Eddy Patuwo, B.; Y. Hu, M. Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting* **1998**, *14*, 35–62. https://doi.org/https://doi.org/10.1016/S0169-2070(97)00044-7.

Jang, J.; Kim, B.; Kim, I. Comparative analysis of deep learning techniques for load forecasting in power systems using single-layer and hybrid models. *Int. Trans. Electr. Energy Syst.* **2024**, *2024*.

Oliveira, H.S.; Oliveira, H.P. Transformers for Energy Forecast. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23156840.

Qiao, L.; Chen, S.; Qu, R.; Ran, R.; Guo, Y.; Tan, W. Electricity consumption prediction based on Transformer-LSTM. In Proceedings of the 2023 3rd International Conference on Electronic Information Engineering and Computer Communication (EIECC), 2023, pp. 228–231. https://doi.org/10.1109/EIECC60864.2023.10456680.

Ahmadipour, M.; Rashedi, E.; Amoozegar, M.a. Electricity Consumption Forecasting Using a Hybrid Approach Based on Transformer Model and LSTM Neural Network. *Iranian Electric Industry Journal of Quality and Productivity* **2024**, *13*, [http://ieijqp.ir/article-1-979-en.docx]. https://doi.org/10.61186/ieijqp.13.2.2.

Pokhrel, S.; Panta, A.K.; Parajuli, S.; Poudel, S.; Adhikari, R.; Guragai, M. IoT based Smart Home with Face Recognition Security. *ICT-Convergence 2022 (ICAEIC-2022)* **2022**.

Qureshi, M.; Arbab, M.A.; Rehman, S.u. Deep learning-based forecasting of electricity consumption. *Sci. Rep.* **2024**, *14*.

Liu, C.; Jin, Z.; Gu, J.; Qiu, C. Short-term load forecasting using a long short-term memory network. In Proceedings of the 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017, pp. 1–6. https://doi.org/10.1109/ISGTEurope.2017.8260110.

Zhong, B. Deep learning integration optimization of electric energy load forecasting and market price based on the ANN–LSTM–transformer method. *Front. Energy Res.* **2023**, *11*.

Saeed, F.; Rehman, A.; Shah, H.A.; Diyan, M.; Chen, J.; Kang, J.M. SmartFormer: Graph-based transformer model for energy load forecasting. *Sustainable Energy Technologies and Assessments* **2025**, *73*, 104133. https://doi.org/https://doi.org/10.1016/j.seta.2024.104133.

Dong, J.; Jiang, Y.; Chen, P.; Li, J.; Wang, Z.; Han, S. Short-term power load forecasting using bidirectional gated recurrent units-based adaptive stacked autoencoder. *International Journal of Electrical Power & Energy Systems* **2025**, *165*, 110459. https://doi.org/https://doi.org/10.1016/j.ijepes.2025.110459.

Aslam, S.; Herodotou, H.; Mohsin, S.M.; Javaid, N.; Ashraf, N.; Aslam, S. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews* **2021**, *144*, None. https://doi.org/10.1016/j.rser.2021.110992.

Shu, J.; Zhang, X.; Yao, Y.; Yi, D.; Gu, B. Graph Spatio-Temporal Attention Network-based Electricity Demand Forecasting. In Proceedings of the 2021 6th International Conference on Power and Renewable Energy (ICPRE), 2021, pp. 792–797. https://doi.org/10.1109/ICPRE52634.2021.9635240.