A GIM-Based Ontological Framework and LSTM-RF Optimization for Real-Time Digital Twin Power Grid Modeling

Cheng Fang, Bo Gao*, Lei Chen, Chu Wei, Qi Luo, Yaqin Lu, Kai Ning
Wuhan Optics Valley Information Technologies Co., Ltd, Hubei Province, Wuhan City, 430075, China
E-mail: gb_master@163.com
*Corresponding Author

Keywords: grid information model, digital twin, power grid model, holographic object encapsulation, long short-term memory

Received: August 10, 2025

Given the problems associated with traditional DTPG (Digital Twin Power Grid) models in large-scale, complex power grid applications, such as the difficulty in fusing multi-source, heterogeneous data and the lack of real-time performance, this study proposes a construction and optimization method with GIM (Grid Information Model) as its core. In the model building stage, this method first realizes the standardized modeling and dynamic state encapsulation of power grid equipment through the extended GIM semantic library and holographic object encapsulation framework, and designs a five-layer dynamic mapping mechanism combined with the Levenshtein distance and BERT (Bidirectional Encoder Representations from Transformers) model to solve the problem of semantic alignment of multi-source data; then develops an event-driven architecture based on Kafka to support real-time data updates and model version backtracking; applies a spatiotemporal dynamic mapping engine, integrates GIS (Geographic Information System) and GIM topological structure, and realizes the spatial positioning and visualization of electrical events. At the optimization level, the self-learning mechanism of the LSTM (Long Short-Term Memory) and RF (Random Forest) combined algorithm is applied to realize the dynamic correction of model parameters and performance improvement. Experiments show that the proposed GIM method takes an average of 4.21 seconds to model, which is much lower than other methods. In the fusion of multi-source heterogeneous data, the field recognition accuracy averages 97.82%. In evaluating model accuracy and adaptability, the GIM method achieves a modeling accuracy of 96.5% and a prediction error rate of only 3.1%. The synchronization delay is stable at 117 to 125 milliseconds. The GIM-based digital twin power grid method in this paper improves the efficiency and accuracy of power grid digital model construction by means of standardized modeling and real-time data fusion, which can quickly respond to changes, accurately locate and warn faults, provide simulation support for dispatching, and enhance power grid reliability and intelligence.

Povzetek: Metoda digitalnega dvojčka, ki temelji na GIM, omogoča hitrejše, natančnejše in realnočasno modeliranje elektroenergetskega omrežja ter izboljšuje zaznavanje napak in zanesljivost sistema.

1 Introduction

The development of smart grid technology is developing rapidly. As the key technology to realize real-time perception and accurate simulation of the state of the power grid, the digital twin power grid is becoming more and more important in the intelligent transformation of the power system.[1], [2], [3], [4]。However, the current digital twin power grid model has frequent problems when dealing with large-scale and complex power grids, and it is difficult to effectively integrate multi-source heterogeneous data, which reduces the modeling accuracy; the dynamic response delay is large, which affects real-time decision-making; traditional modeling methods cannot effectively characterize the topological characteristics of the power grid, and the calculation efficiency is low and the model extensibility is poor when processing massive amounts of real-time data.[5], [6], [7].Traditional CIM models use planar topology to describe the power grid, which is difficult to accurately capture the spatial structure and dynamic coupling relationship. Subjective deviations and semantic ambiguities are easily introduced, and the manual modeling efficiency is low and the modeling error is high. Although techniques like BIM and KG have clear structural benefits, they are challenging to meet in real time due to their high synchronization delays and high computing resource consumption. Based on this, this study suggests three questions to confirm the efficacy of the method: can the synchronization delay be stabilized within 150 milliseconds in real-time; can the field recognition accuracy be increased to over 95% and semantic ambiguity decreased; and can the prediction error rate of key operating parameters be decreased to less than 5% in terms of dynamic optimization ability.

This work creatively suggests a GIM-centric strategy to address fundamental issues in digital twin power grid modeling. By utilizing topological modeling capabilities and standardized semantic representation, GIM provides the fundamental basis for building high-precision digital twin systems. We created a GIM-based holographic object encapsulation framework during model development to facilitate dynamic state encapsulation and uniform equipment modeling. A multi-source heterogeneous data mapping system was also created to improve the automation and accuracy of data fusion. An LSTM-RF hybrid approach was developed for model improvement in order to create a parameter self-learning mechanism, which greatly increased prediction accuracy and dynamic adaptability. High modeling accuracy and minimal synchronization delay are demonstrated by experimental results, offering a practical way to implement digital twin power grids widely.

This study enhances the creation and optimization of DTPG models. In order to address complicated grid topologies and multi-source heterogeneous data fusion, it suggests a DTPG model construction framework based on GIM. For such data, a GIM dynamic mapping technique is developed, improving automation and fusion accuracy. Additionally, it creatively builds a self-learning parameter optimization system for real-time key parameter pattern recognition and model correction using the LSTM-RF method. These techniques provide great support for the intelligent development of power systems and the widespread application of the DTPG paradigm.

2 Related work

Existing research on DTPG models mainly centers on multi-source data integration, unified modeling frameworks, and real-time mapping [8], [9], [10]. Technical challenges vary: Jiang Zongmin's OKDD model [11] achieves device-level info unification via CIM extension but neglects grid topology dynamics, leading to a 12.7% modeling error in large-scale scenarios. Juntao et al.'s platform [12] integrates station-side data for grid O&M, while Jing et al.'s topology mapping [13] boosts device mapping efficiency to 50% but with a 420ms delay, failing real-time needs. Liu Ge's relay protection twin model [14] cuts configuration time by 35% but lacks dynamic optimization, requiring reconstruction for topology changes. Two major gaps persist: unresolved real-time model updates under dynamic topology and the absence of a unified framework integrating structural perception and dynamic optimization.

In recent years, GIM has been applied in power grid modeling to better capture spatial structures and complex node relationships [15], [16], [17]. Traditional CIM models, with planar topology, offer less than 65% time-series feature resolution for smart substation secondary circuits. Zhang et al. [18] developed a 3D GIM-based visualization system, improving spatial relationship recognition accuracy to 91.3%. Rong et al.'s framework [19] reduced UHV project modeling costs by 22%, but its static indicator system can't support real-time optimization. Shi et al.'s GIM solution [20] enhanced design change response speed, and Zhang et al.'s platform [21] reduced cross-span errors, proving GIM's superiority in spatial modeling. Unlike CIM's static topology, GIM dynamically resolves time-varying node coupling via object-oriented encapsulation and relational graphs, providing a structural basis for real-time optimization. This paper proposes a GIM-based digital twin grid model with dynamic topology perception via node relationship graphs, integrating the BERT algorithm for semantic alignment in multi-source data, effectively addressing limitations in real-time responsiveness and optimization efficiency.

In order to present the advantages of GIM-based digital twin power grid model, the key performance indicators of the methods involved in relevant literature are summarized in Table 1.

Table 1: Comparison of key performance of digital twin power grid modeling methods

Performance indicators	GIM method	OKDD [10]	Intelligent digital twin operation and maintenance platform [11]	Topology mapping framework [12]
Modeling error/accuracy	Modeling accuracy 96.5%	Modeling error 12.7%	-	-
Calculation/synchronization delay (milliseconds)	Synchronization delay 117-125	-	-	Calculation delay 420
Efficiency improvement (such as shorter configuration time)	Modeling time 4.21 seconds	-	-	Device mapping efficiency increased to 50%
Multi-source data fusion capability	Field recognition accuracy rate 97.82%	-	Realize the deep integration of station-side data	-

3 Twin power grid model construction system based on GIM

In order to clearly present the overall technical architecture and data flow logic of the GIM-based digital twin power grid model, this paper constructs the system architecture diagram as shown in Figure 1. The architecture follows the hierarchical design concept from physical perception to intelligent application.

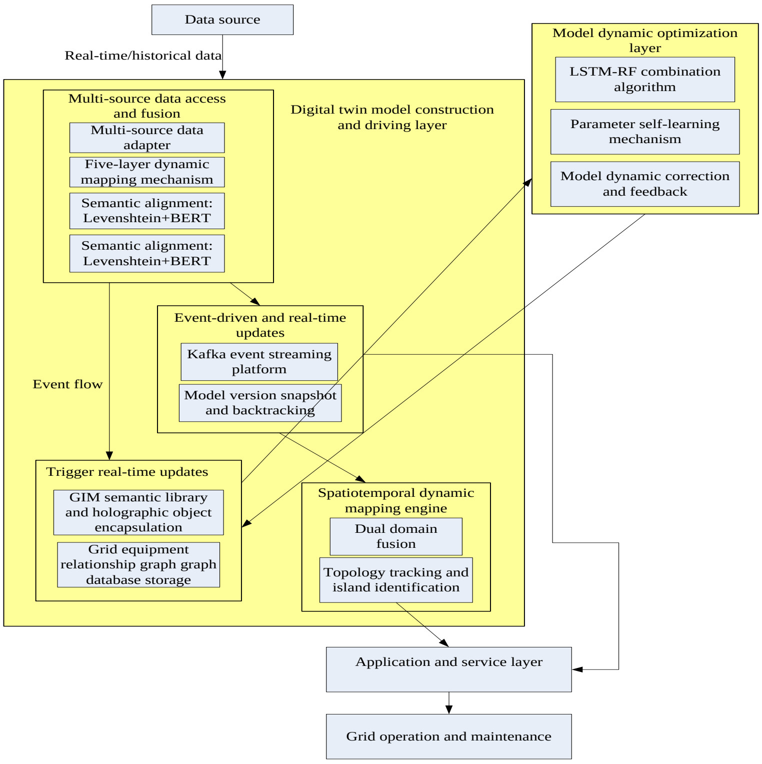

Figure 1: Framework diagram of twin power grid model of GIM

3.1 GIM semantic extension and holographic object encapsulation

In digital twin grid construction, semantic consistency and complete expression of models are crucial for high-precision virtual-real mapping. Traditional power grid information modeling suffers from issues such as data fragmentation, semantic redundancy, and inconsistent structures, which hinder digital transformation. Based on the IEC 61970/61968 standard and practical requirements, this paper constructs an extensible GIM semantic library and proposes a holographic object encapsulation framework to achieve standardized modeling and consistent semantic management of all elements [22], [23], [24]. To ensure semantic mapping consistency across multiple systems, the GIM semantic library employs ontological logic and standard languages like RDFS (Resource Description Framework Schema) to build domain semantic relationship graphs. The procedure entails consistently encapsulating pertinent relationship classes and expanding attributes to construct equipment meta-class collections after first abstracting power grid component ontologies to clarify attributes and logical relationships. By building relationship graphs using ontological logic, the GIM semantic library improves model consistency, inferential qualities, and structural extensibility. Format for defining universal equipment objects:

$$E_i=\langle T_i,A_i,R_i,S_i\rangle \tag{1}$$

E_i represents the i-th power grid device; T_i represents the device type.

In order to provide a knowledge graph storage structure that offers data support for model semantic reasoning and structural recognition, the GIM semantic library incorporates more than 300 attribute fields and more than 100 common equipment classes based on this semantic model. It does this by using a graph database. This study suggests a holographic object encapsulation (HOE) framework that uniformly wraps device shape, electrical properties, asset information, and state parameters into computable model objects in order to accommodate the dynamic interaction requirements of DT situations. The framework uses a very unified and low-coupling encapsulation class design to accomplish the natural integration of static and dynamic properties. The following are the definitions:

$$HOE_i=\langle F_i,Q_i,B_i,D_i\rangle \tag{2}$$

F_i represents the geometric structure model of the equipment; Q_i represents its physical electrical parameters; B_i represents its asset management information; and D_i represents its current operating status set.

The API offers services that support a variety of actions, including model addition, deletion, modification, querying, topology changes, and state refreshes, after the integration of the HOE framework with the GIM semantic library. It incorporates permission and version control for model security and traceability, and it is based on binding unique device identification codes to allow granular data interaction for individual devices. With relevant constraint links built, the model facilitates state logic constraint reasoning during ontology encapsulation, such as setting the flow of a connected branch to 0 when the circuit breaker is disconnected.

The generation of equipment type and attribute during the creation of the enhanced GIM semantic repository closely follows ontology engineering approaches. The approach starts with a methodical examination of the fundamental skills in power grid areas. This guarantees the usefulness and completeness of the ontology by formulating and answering "capability assessment questions"—such as equipment status changes during fault isolation and tracking power supply sources/pathways in crucial business scenarios. The IEC 61970/61968 standard common model is dissected, building upon this base. Concepts are retrieved and attributes are generalized by combining a top-down and bottom-up methodology with field documentation. While the bottom-up strategy reverse engineers some data sources to extract actual data fields and translate them to ontology, the top-down method uses business processes to define equipment types and key status attributes.

Several validation procedures are used to guarantee the quality of the built ontology. Formal verification prevents inconsistencies in class definitions and relationships by using axioms and rules from the Resource Description Framework Schema (RDFS) and Web Ontology Language (OWL) to check logical consistency. This method's capacity to provide clear responses to "capability assessment questions" like power supply path tracing and affected equipment identification is demonstrated by empirical verification, which applies it to real substation model construction scenarios. In order to ensure that the more than 300 attribute syntaxes and more than 100 equipment classes that are eventually created are logically and semantically accurate in capturing grid expertise and operational rules, this methodology incorporates capability requirements throughout the entire ontology construction and validation process. This gives digital twin models a strong, inferable semantic basis.

3.2 Five-layer dynamic mapping mechanism for multi-source heterogeneous data

When building a DTPG model based on GIM, it is important to efficiently and accurately access and integrate heterogeneous power data. Typical data sources, such as SCADA (Supervisory Control and Data Acquisition) systems and GIS platforms, exhibit significant differences in multiple dimensions, including data structure and semantic naming, which lead to serious defects in traditional modeling methods in terms of automation and consistency control. To build a panoramic twin power grid, a high-performance access mechanism is necessary, as well as unified modeling at the semantic layer.

This study proposes a five-layer multi-source data access and mapping mechanism. In view of the structural differences of different data sources, an extensible data adapter framework is designed. For structured data, a field parsing and mapping strategy is adopted; for semi-structured or unstructured data, a template-driven data extraction mechanism is used, and then parsed into a unified intermediate representation format through the ETL (Extract, Transform, Load) process. The high-frequency telemetry of the SCADA system is accessed in real-time through the OPC (Open Platform Communications) interface and converted into a structured JSON (JavaScript Object Notation) object. The unified intermediate format lays the foundation for subsequent data fusion and unified modeling. The expression is:

$$D_{std}=\langle T,H,\{k_1:k_2,k_1:k_2,\cdots k_i\}\rangle \tag{3}$$

T represents the timestamp; H represents the data source label; $\{k_1:k_i\}$ is the attribute key-value pair set.

After the structure is unified, the model builds a field mapping table based on the extended GIM semantic model to achieve semantic alignment between multi-source data and the target model. The model first constructs a multi-layer matching dictionary based on a historical corpus, and then utilizes regular matching, a rule engine, combined with Levenshtein distance and the BERT model to generate a similarity matrix, thereby determining the optimal matching path. The matching algorithm is simplified as follows:

$$M(a,b)=\alpha*Sim_{str}(\mathrm{h}_a,\mathrm{h}_b)+\beta*Sim_{sem}(\mathrm{h}_a,\mathrm{h}_b) \tag{4}$$

h_a and h_b are the original field and GIM field, respectively; Sim_{str} is the string similarity score; Sim_{sem} is the semantic similarity score based on the word vector model; and α and β are the weight coefficients. Finally, the field pair with the largest $M(a,b)$ is selected as the mapping result. Equation (6) In the semantic alignment model, the weight coefficients α and β are not dynamically obtained by machine learning, but are pre-set as static hyperparameters based on a priori analysis of the relative importance of string and semantic matching.After grid search combined with verification set performance orthogonalization tuning, it is determined that the optimal combination is α=0.3 and β=0.7, indicating that semantic similarity plays a leading role in grid field alignment, and string similarity plays an auxiliary correction role.The performance of the model is sensitive to the weight configuration. When β changes in the range of 0.6 to 0.8, the alignment accuracy rate is relatively stable, and the accuracy rate decreases significantly when β<0.5, reflecting the irreplaceability of semantic information in solving the problem of synonymous and heteromorphic words.Completely relying on any one dimension will degrade performance. This sensitivity analysis confirms the necessity of dual-path design, and fine-tuning the weights within a reasonable range will not cause severe performance fluctuations, enhancing engineering practicality.The semantic similarity heat map of power system parameter fields is shown in Figure 2.

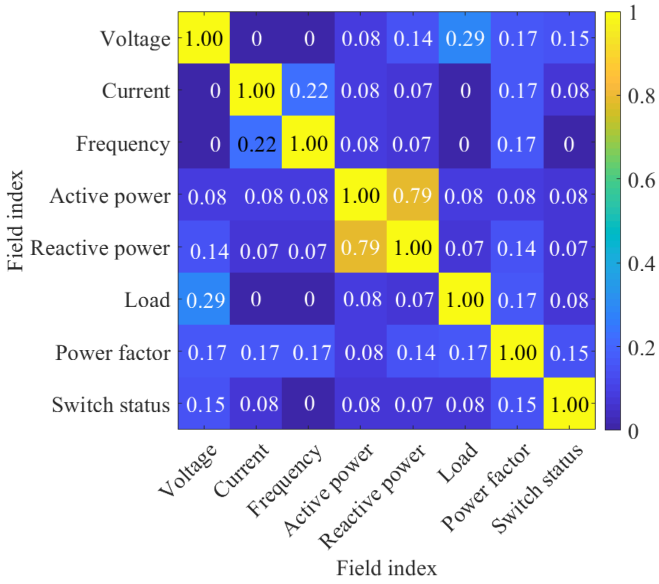

Figure 2: Heat map of semantic similarity of power grid model parameter fields

Figure 2 shows the semantic similarity between power grid model parameter fields. First, new field names are defined to represent common measurement data types (such as voltage and current). A matrix is then initialized to store similarity values between fields. Semantic similarity is quantified by calculating the Levenshtein distance between each pair of fields and converting this into a score-filled matrix. Heatmap cells display specific values with adaptive color for easier reading.

An event-driven mechanism based on the Kafka stream processing architecture is used to guarantee that the equipment's operation state is in line with the DT model in real-time, supporting dynamic model changes and real-time responses [25]. As a Kafka message producer, the integrated data adapter module encapsulates semantically standardized and structurally altered grid operation data in an intermediate JSON format before forwarding it to a specified topic. Numerous real-time events are reliably cached and distributed sequentially using Kafka. As a consumer, the backend GIM service module monitors message streams, subscribes to topics, and initiates processing. Additionally, the model predefines three fundamental event categories: state mutation events, parameter change events, and device commissioning and decommissioning events.

We carefully planned the fundamental technological specifications and fault tolerance methods when creating the Kafka-based event-driven architecture. Based on benchmark stress testing and realistic business scenario load modeling, the system is designed to successfully process 10,000–15,000 device status signals per second. In order to accomplish this, we used a composite partition key method that divides messages equally among 20 Kafka partitions according to device type IDs and physical location. Robust performance is ensured by this setup in conjunction with multi-consumer instance groups for parallel processing and horizontal scaling. The monitoring module monitors partition consumer latency when event backpressure is encountered. The resource scheduler adds new consumer instances through container orchestration platforms to initiate horizontal scaling when thresholds are surpassed.

We design a monotonically growing logical timestamp that embeds message metadata while decoupling from physical clocks while maintaining causality in order to address message out-of-order problems brought on by network jitter in power grid scenarios. In addition to rearranging messages according to device granularity and using Bloom filters to identify and eliminate duplicate messages, the consumer side implements a time-window-based sorting buffer technique that defaults to a 500-millisecond window. This method reduces processing delay while guaranteeing that status updates are made in the proper chronological order.Given that power grid digital twin models place a higher priority on data integrity than latency sensitivity, the system uses a "at least once" delivery semantics to ensure consistency. Through the use of database uniqueness constraints and Kafka producer idempotence features to avoid message duplication, as well as periodic asynchronous snapshots and state checkpoints, the system allows users to recover from recent consensus states following system restarts, ultimately leading to consistency.

Important details like the device's unique identification are contained in event messages. Following feature extraction, the model uses the device ID to find the matching node in the GIM model. To find out if the target device has undergone any significant modifications, the most recent snapshot version is then requested. To make sure the twin is compatible with the field state, the model fields or topology are automatically updated if such is the case. Following the modification, the system writes the operational data to the version management module and bundles it into a version snapshot. This snapshot creates a continuous evolution trajectory by comparing the model state before and after the change.The system uses a data-driven model repair method to automatically enhance the model when combining multiple sources. For instance, the system initiates a comparison operation to confirm whether the geographical coordinates match the logical position if the location of the GIS device differs from the SCADA status [26], [27], [28]. It automatically suggests a location update and asks the engineer to approve it if the offset is greater than the threshold. The coordinate consistency verification formula that follows is applied:

$$d_{pos}=\sqrt{(xSCADA\text{-}xGIS)^2+(YSCADA\text{-}YGIS)^2} \tag{5}$$

When $d_{pos}>d_{THRESH}$, it is considered a spatial semantic conflict, triggering the model correction logic.

With its unified architecture that combines dynamic perception, semantic intelligence, and real-time response, GIM exhibits impressive technical advantages. First, GIM builds a knowledge network with dynamic ontology reasoning capabilities by utilizing ontological logic and an expanded semantic library. By defining device static properties and enhancing automatic semantic verification through integrated logical rules, this innovation solves the drawbacks of standard CIM's static flat description, guaranteeing semantic consistency and lowering ambiguity. Second, a strong rule engine built into GIM automates tasks like outlier detection and unit standardization. To guarantee physically accurate data fusion and prevent the logical errors that come with purely data-driven methods, the system uses strict rules based on physical principles and logical limitations. These three synergistic advancements enable GIM to surpass existing methods in multiple dimensions.

Pseudo-code of the five-layer dynamic mapping mechanism:

Input: Raw data records D_{raw}, GIM semantic library G

Output: Standardized GIM data objects D_{std}

1. Structure parsing
   - -Parse raw data fields: F ← $\text{parseFields}(D_{raw})$
   - -Add data source tags: F ←∪ F{source: $D_{raw}.\text{source}$}
2. Semantic mapping
   - for f ∈ F do
   - for g ∈ G do
     - -Calculate comprehensive similarity : $\mathrm{M(f,g) = \alpha * Levenshtein(f,g) + \beta * BERT(f,g)}$
   - end for
   - -Choose the best match : $g_{best} \leftarrow arg\ max_g M(f,g)$
   - if $M(f,g_{best}) > \theta then$
   - -Establish mapping : $f \rightarrow g_{best}$
   - end if
   - end for
3. Unit normalization
   - -for each mapped field (f,g) do
     - if $f.unit \neq g.unit$ then
       - -Perform unit conversion : $f.value \leftarrow f.value * getFactor(f.unit, g.unit)$
     - end if
   - end for
4. Quality verification
   - -for each numeric field f do
     - -If $f.value$ exceeds the effective range then
       - -Use the sliding window to correct the mean
     - end if
   - end for

5.Event Publishing
   - -Building Standard Objects: $D_{std} \leftarrow bulidStandarObject(F)$
   - -Publishing Messages:$Kafka.produce(topic, D_{std})$

Return : D_{std}

3.3 Dynamic spatiotemporal mapping and event-driven architecture

To enable DTPG models to play a role in business scenarios such as operation monitoring, they need to be abstracted from the underlying model layer into an upper-layer, interactive, and callable service form, and utilize multi-dimensional visualization to intuitively display the power grid status, etc. [29], [30]. This paper constructs a functional system with twin service encapsulation as its core. A spatiotemporal dynamic mapping engine and a microservice interface support the real-time mapping of multi-dimensional power grid scenarios, enabling the transition from data twins to cognitive twins.

At the spatial level, twin visualization requires both static representation of the physical structure of the power grid and dynamic rendering capabilities. This is especially true for real-time scenarios like fault location, where electrical logic and geographic information must be synchronized. This paper develops a spatiotemporal dynamic mapping engine that integrates GIS basemaps with GIM topology to achieve dual-domain convergence. At the device level, each entity object has a unique identifier Z and geographic coordinates (x_i,y_i) in the GIM, and the electrical connection relationship is represented by the adjacency matrix Q_{ij}. By constructing the electrical connectivity graph $G=(V,E)$, the electrical connections are mapped to the GIS spatial path to achieve logical relationship visualization. Based on a topology tracking algorithm, the system can locate the abnormal path in real-time and display its spatial direction on a GIS map after an electrical event occurs. If the state of a node v_i suddenly changes, the system uses a depth-first search (DFS) to find the reachable area $R_i \subset V$ in the topology map and plots the spatial range based on the node coordinate set, providing powerful support for grid operation monitoring and fault handling.

In actual operation, to accurately depict the impact area of electrical events, the system performs dynamic path tracing and island identification based on the electrical diagram topology. When a device's status suddenly changes, a DFS search for the reachable domain is performed starting from that node. The edge weight function is then used to determine the power supply and switchable paths, thereby determining the fault impact area. At the same time, to achieve real-time rendering of equipment operating parameters, a power flow solver module is applied as a callable service. It can quickly calculate and visualize the electrical parameters of each node in the power grid under different operating conditions and solve the following node power balance equation:

$$Q_i=\sum_{j=1}^{m}|V_i||V_j|\left(G_{ij}cos\theta_{ij}+A_{ij}sin\theta_{ij}\right) \tag{6}$$

$$P_i=\sum_{j=1}^{m}|V_i||V_j|\left(G_{ij}sin\theta_{ij}+A_{ij}cos\theta_{ij}\right) \tag{7}$$

Q_i and P_i represent the active and reactive power injected into node i, respectively; $|V_i|$ is the voltage amplitude; $\theta_{ij}=\theta_i-\theta_j$ is the voltage phase angle difference; G_{ij} and A_{ij} are the real and imaginary parts of the admittance.In terms of topology analysis services, the model supports multiple modes, including path finding, island identification, and power supply range calculation. For example, given a busbar node, it traverses its connected branches and recursively expands the set of adjacent nodes to calculate the power supply capacity boundary. Island identification is based on connected domain partitioning, decomposing the electrical diagram into a set of subgraphs to determine whether there are isolated areas without external branch connections [31], [32], [33].

4 Model dynamic optimization method based on LSTM-RF

4.1 Parameter self-learning mechanism design

A DTPG model is a real-time mirror of the physical grid system. The accurate modeling of electrical parameters and operating circumstances is essential to the simulation's accuracy. Static parameter choices, however, may differ from real-world conditions due to various factors such load fluctuations, line impedance drift, and equipment aging, which lowers the precision of model simulations and forecasts. This research uses an LSTM-RF combination algorithm-based self-learning optimization technique to allow the twin model to automatically adapt when the grid changes. This technique dynamically identifies important patterns of parameter variation by performing time series modeling and feature extraction using past monitoring data. High-precision self-evolution of the grid twin is made possible by its real-time adjustment of GIM settings through model feedback.

This study's dynamic optimization approach for the Random Forest (RF) model mixes real-time inference with periodic retraining. Historical data covering roughly 24 hours is taken out of the time series database at the beginning of each cycle. In order to create new training samples for complete retraining of the RF model, the LSTM network first extracts temporal features from this data, which are subsequently merged with matching real parameter labels. Following training, the cycle's real-time data is fed into the modified RF model's inference parameters via the same LSTM pipeline until the subsequent cycle starts. Through fixed-cycle batch processing, as shown in Figure 3, this approach allows the model to adjust to grid fluctuations while preserving a balance between computational costs and efficiency.

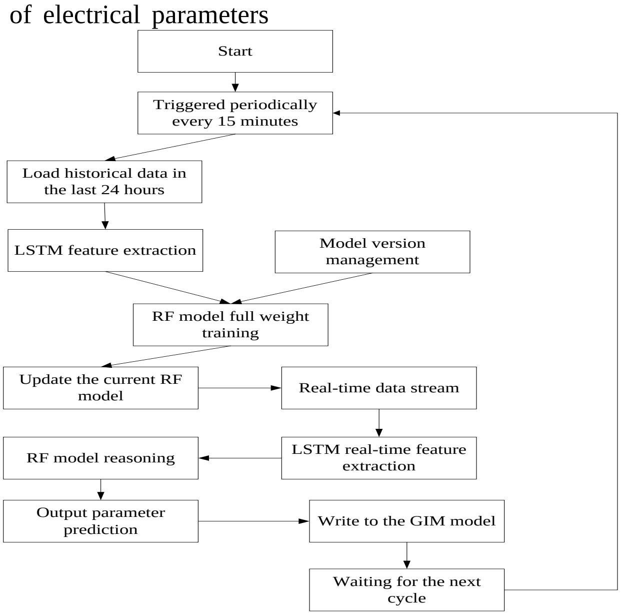

Figure 3: Model flow chart of LSTM-RF

Through periodic model retraining, the procedure depicted in Figure 3 allows the system to adjust to dynamic changes in the power grid, satisfying the low-latency criteria for real-time business inference. When performance deteriorates, the model version management system allows for a quick rollback to the prior stable version by maintaining RF model snapshots.

Building a learning framework that automatically derives patterns of parameter evolution from prior operational data is the fundamental component of the LSTM-RF-GIM technique.The LSTM network is responsible for modeling the long-term dependency structure of time series and extracting the dynamic evolution trends of variables such as voltage under different operating conditions [34], [35], [36]. RF, a nonlinear ensemble learning model, combines feature importance assessment to interpretably model the role of each parameter in prediction error, improving the model's robustness and noise resistance. Assuming that at a certain device level, given the operating parameters and observed state sequence, the goal is to construct a function that approximates the optimal parameter estimates and reflects the device's true physical response capability. After modeling, the LSTM network outputs a sequence of hidden states, which is then input into the RF model as a high-dimensional time feature. Further extraction of the factor combination most sensitive to changes in the target parameter is performed. Assuming that a_t=g_t, then the RF model learning function is:

$$\hat{\theta}_i(t)=RF_i(a_t),\quad \forall i\in[1,m] \tag{8}$$

RF consists of multiple decision trees trained on different features and subsets of samples. Its output is the average of the predictions from each tree. It has good generalization capabilities and is suitable for high-dimensional multivariate modeling. The application of a feature importance assessment mechanism can identify physical state variables that drive target parameter drift, thereby enhancing the model's explanatory power and facilitating predictive maintenance of equipment. After the prediction is completed, the GIM model parameter fields are dynamically updated according to a specific update mechanism.

$$\theta_i^{new}(t)=\alpha*\hat{\theta}_i(t)+(1-\alpha)*\theta_i^{prev}(t) \tag{9}$$

$\alpha\in[0,1]$ is the updated weight coefficient; $\theta_i^{prev}(t)$ is the historical parameter value; $\theta_i^{new}(t)$ is the current predicted value. The updated parameters replace the corresponding fields of the device object in the GIM model, enabling global dynamic convergence of the twin.

In this study's LSTM-RF hybrid architecture, the LSTM network is tasked with modeling complete temporal sequences and extracting physically meaningful high-level features from raw time-series data, rather than performing end-to-end predictions. The network is subjected to multi-layer temporal processing when operational status sequences are input, and the final hidden state output functions as a condensed feature vector. Within the observation frame, this vector captures long-term dependencies, trends, and dynamic patterns. Before being entered into the ensemble of decision trees, the vector is concatenated with static features such as the device model and ambient temperature in the RF model for nonlinear regression analysis. The RF model uses ensemble learning to improve generalization capabilities by learning intricate mappings from joint features to target parameters.RF's advantages in processing high-dimensional data and importance assessment are combined with LSTM's ability to capture temporal dynamics in the "LSTM feature extraction + RF regression" pipeline design. The method offers a dependable route for accurate parameter estimation in power grids by separating temporal modeling from feature reasoning. This allows the system to integrate deep network representation capabilities with the stability and interpretability of ensemble learning in small sample circumstances.

4.2 Optimizing closed-loop implementation

Using a sliding window approach, the model is trained incrementally at regular intervals, like every 15 minutes, by extracting the most recent time frame from previous data. To avoid overfitting, early halting and dropout techniques are also used. The model outputs are promptly verified for consistency both before and after parameter modifications. To guarantee steady evolution, the model is rolled back whenever the error rises above a certain point. By combining time series modeling and nonlinear regression analysis, a self-learning optimization technique based on LSTM-RF-GIM dynamically determines and iteratively updates GIM model parameters. The algorithm layer guarantees consistent and comprehensible parameter predictions, while the model layer is able to detect changes in the operational state. This offers a workable way to optimize DTPG models while they are in use.

Real-time feedback and model changes are coordinated via a hierarchical processing mechanism in the system architecture. Response times in tasks like data acquisition and state inference are measured using real-time performance indicators (117–125 ms delay), which use a pre-trained LSTM-RF model for rapid inference without waiting for updates. An asynchronous background operation retrains the model to optimize parameters every 15 minutes. In order to balance accuracy and real-time performance, the system tracks discrepancies between physical measurements and inferences made in real-time. When the average absolute error over a certain threshold, it automatically updates the model. However, it also has a periodic mechanism to prevent excessive updates.Key LSTM-RF hyperparameters, including the LSTM's double-layer structure with 128 hidden units per layer, 0.2 dropout rate, Adam optimizer (learning rate 0.001), and 60 historical time steps, were adjusted manually and by grid search. With mean square error as the splitting criterion, the random forest employs 500 decision trees (max depth 25, min node samples 5), randomly selecting 328 characteristics per tree. In order to achieve the best bias-variance tradeoff on the validation set, the model trains for 100 rounds with early halting if validation loss does not improve for 10 rounds in a row.

The GIM-based digital twin power grid model employs a hierarchical computational complexity structure. During data ingestion, the Kafka streaming platform handles tasks with O(N) time complexity (where N represents message count), achieving linear throughput growth with horizontal scaling. Multi-source data mapping constitutes the system bottleneck. Semantic alignment operations involve Levenshtein distance calculations with O(m×n) complexity dependent on string lengths m and n, while BERT semantic similarity computations utilize $O(L^2)$ complexity tied to sequence length L. However, pre-trained models and batch processing can reduce per-operation costs. The GIM modeling phase leverages graph databases with O(|V| + |E|) complexity. The LSTM-RF inference pipeline features LSTM forward propagation complexity $O(t*h^2)$ related to sequence length t and hidden layer dimension h, alongside RF inference complexity O(T*log(D)), which correlates with decision tree quantity T and feature dimension D. The system achieves near real-time performance through pipeline architecture and parallel computing, though data mapping and LSTM feature extraction remain resource-intensive processes.

An LSTM-RF hybrid approach is used in this study to optimize the model while taking engineering applications, modeling requirements, and power grid data characteristics (noise, nonlinearity, and temporal dependency) into account. The LSTM network is appropriate for periodic and trend-based power system data because it successfully captures long-term temporal correlations in parameters like voltage and load, resolving classic RNN gradient training concerns for lengthy sequences. As an ensemble model, RF facilitates the interpretation of complex events by rapidly processing high-dimensional features and providing interpretability through feature importance evaluation. The LSTM-RF combination meets the requirements of the digital twin system and performs better in terms of transparency than BiGRU and TCN.But there are restrictions: RF exhibits instability with extreme anomalies and limited adaptation to rapid data shifts; excessive hyperparameters hinder deployment; and LSTM's sequential computing limits efficiency for ultra-long sequences. Future research aims to improve practicality and robustness in industrial contexts, use online learning for data drift adaptation, and incorporate attention techniques to concentrate on important time points.

5 Performance evaluation and analysis of DT grid models

5.1 Introduction to the experimental environment and dataset

To comprehensively evaluate the construction and optimization method of the DTPG model based on GIM, the experiment employs a multi-source data fusion scheme, with the experimental parameters listed in Table 2.

Table 2: Experimental environment configuration and tool parameters

Category	Detailed information
CPU(Central Processing Unit)	Intel Core i9-10900K @ 3.70GHz
Memory	64GB
GPU(Graphics Processing Unit)	NVIDIA RTX 3080
Operating system	Windows 10 Pro
Programming language	Python 3.8, MATLAB R2021b
Graph database	Neo4j 4.4
Deep learning framework	PyTorch 1.9
Visualization tools	Matplotlib, Seaborn, Plotly
Data flow processing	Apache Kafka

The hardware used is an Intel Core i9-10900K processor, 64GB of RAM, and an NVIDIA RTX 3080 to power data processing and graphics computing. Windows 10 Pro is used as the operating system, Python 3.8 is used for programming, and PyTorch 1.9 is used for training.

The data set of this research is taken from the actual operating system of a power saving company from January 2022 to June 2023, covering about 1.5 billion multi-source heterogeneous data records, covering 128 major substations and related transmission and distribution networks in the province.The data includes about 1.2 billion second-level telemetry and telemetry data from the SCADA system, about 180 million synchronous phasor data from the PMU (Phasor Measurement Unit), about 12 million spatial data from the GIS platform, and about 1 million static parameter data from the equipment asset management system.During preprocessing, first check and clean the original data, remove obvious error records, and use the spatiotemporal correlation interpolation method to repair the missing values, including rule verification based on the operating characteristics of the equipment, linear interpolation of adjacent time periods, and collaborative compensation of data from spatially adjacent sites.For the problem of timestamp alignment of different data sources, the SCADA data is used as the reference and the sliding time window is used to synchronize. The window size is set to 100 milliseconds to 1 second according to the data type.The final standardized data set contains 85 operating parameters of key equipment types and 328 core measurement points, which provides sufficient data support for model construction and verification.The real operation data set of a power-saving company used in this study has obvious advantages.It combines SCADA conventional telemetry and telemetry data with PMU high sampling frequency synchronous phasor data, with high time resolution, and can accurately capture the transient and dynamic response of the power grid, laying the foundation for the high-fidelity simulation of the digital twin.The data set covers multiple condition information of key equipment, with wide space and complete dimensions, which makes up for the shortcomings of the existing data set in terms of scale, complexity and authenticity, and makes the model training and evaluation more representative and engineering value.

5.2 Model building performance

When building DTPG models, modeling consistency and efficiency are crucial to model reliability and scalability. To evaluate the semantic standardization and efficiency of different modeling methods, this paper conducts an experimental comparison of five mainstream approaches: GIM, CIM, Manual Modeling (MM), Building Information Modeling (BIM), and Knowledge Graph (KG). The experiment focuses on the number of semantic ambiguities and modeling time, simulating the performance of each method in a standard modeling task. Ten experiments are conducted, with the results shown in Figure 4.

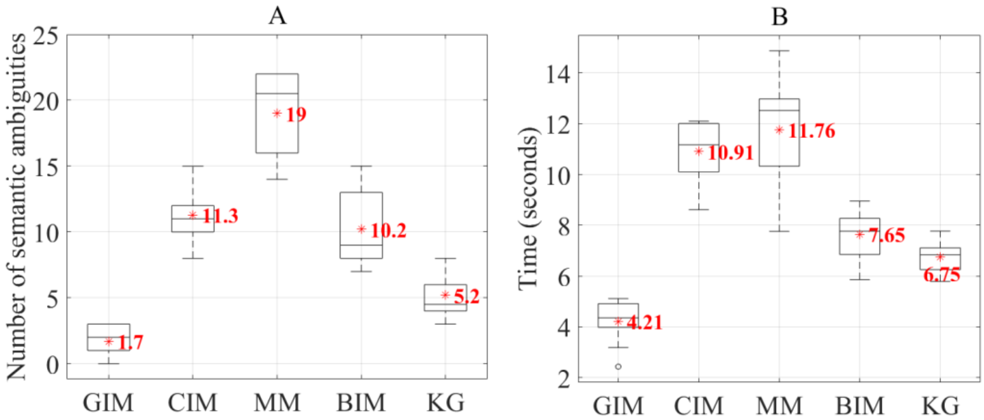

Figure 4: Comparison of semantic ambiguity and modeling time for different modeling methods; Figure 4A: Number of semantic ambiguities, Figure 4B: Modeling time (seconds),

The suggested approach is contrasted with alternative models in Figure 4 with respect to modeling time and semantic ambiguity. With an average of just 1.7 semantic ambiguities—much fewer than those of the other methods—the GIM approach performs noticeably better than the others in terms of semantic ambiguity, as seen in Figure 4A. The average semantic ambiguity of the CIM approach is 11.3 despite significant semantic standardization. The MM approach uses manual labor, which has an average semantic ambiguity of 19 and is subject to subjective bias and mistakes. With average semantic ambiguities of 10.2 and 5.2, respectively, the BIM and KG approaches—which make use of structured data and knowledge representation techniques—remain below the GIM method.As shown in Figure 4B, the GIM method also performs well in terms of modeling time, with an average modeling time of 4.21 seconds, which is significantly shorter than that of other methods. This is because of its well-designed data processing and model-building features. Large-scale power grid data can be processed quickly thanks to sophisticated algorithms and tools, which also make model training and optimization more effective. The average modeling time for the CIM approach is 10.91 seconds. The MM approach is the least effective since it necessitates a great deal of human labor and has an average modeling time of 11.76 seconds. Despite using automation technology, the BIM and KG approaches fall short of the GIM method with average modeling times of 7.65 and 6.75 seconds, respectively. This implies that the GIM approach enhances the DTPG models' accuracy, scalability, practicality, and dependability by decreasing semantic ambiguity and cutting down on modeling time, proving its broad applicability and promise for adoption.

Disparities in field semantics and uneven unit standards decrease data quality and application efficacy when merging heterogeneous data from several sources. In order to assess their advantages and disadvantages in managing complex, diverse data semantics, this study uses the GIM dynamic mapping mechanism in comparison to knowledge graphs (KG), Random Forest (RF), static mapping processing (SMP), and BERT. RF and BERT were evaluated separately as baselines and were not incorporated into the GIM framework: Without rule engines or GIM domain knowledge, BERT used its pre-trained language model to determine the semantic similarity of grid fields; RF used statistical feature selection and conventional feature engineering with bag-of-words models for field classification.The general benefits of the GIM framework are objectively highlighted by this configuration. Unit homogeneity and field recognition accuracy are the two main measures that are the focus of the experiment. Figure 5 displays the findings.

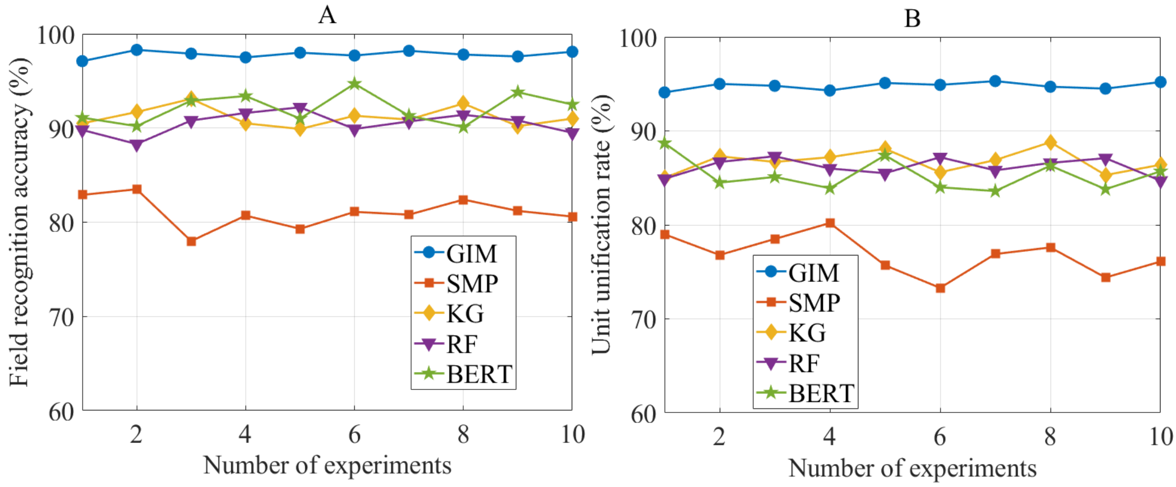

Figure 5: Performance comparison of multi-source data dynamic mapping methods; Figure 5A: Field Recognition Accuracy, Figure 5B: Unit Consistency

As shown in Figure 5, the GIM dynamic mapping mechanism outperforms other methods overall, demonstrating superior flexibility and semantic processing. As shown in Figure 5A, the GIM technique achieves an average field recognition accuracy of 97.82%, which is much higher than SMP (81.05%), KG (91.17%), RF (90.5%), and BERT (92.1%). In all 10 studies, GIM consistently performs well, with a minimum of 97.1% and a high of 98.3%. SMP's strict semantic reasoning causes significant accuracy swings, with a minimum accuracy of 78%. KG improves field matching capabilities based on a structured knowledge base, but due to predetermined rules and knowledge coverage limitations, its accuracy is lower than GIM's.With an average unit consistency rate of 94.79%, as illustrated in Figure 5B, GIM outperforms SMP (76.85%), KG (86.73%), RF (86.18%), and BERT (85.3%). In the sixth trial, for instance, GIM attains a unit consistency rate of 94.9% while SMP only manages 73.3%.

As power grid intelligence and digitization continue to advance, the creation of DTPG models has become an essential technology for achieving efficient grid operation. The different modeling approaches differ greatly in terms of data fusion, semantic expression, and dynamic adaptability. Using a variety of performance metrics, including model correctness and adaptability, this experiment compares the GIM technique with DT models based on BIM, KG, RF, and LSTM, as shown in Table 3.

Table 3: Comprehensive performance comparison of different DTPG modeling methods

Indicator	GIM	BIM	KG	RF	LSTM
Modeling accuracy (%)	96.5	88.2	91.7	92.5	90.3
Prediction error rate (%)	3.1	7.6	5.6	4.8	4.3
Model update frequency	10	5	6	8	7
Computing resource consumption (%)	35	63	51	45	42
Adaptability score (1-10)	9.2	7.8	8.1	8.4	8.6
Semantic consistency (%)	98.4	85.6	89.5	92.7	93.8
Parameter dynamic correction ability (1-10)	9.5	7	8.3	8.5	8.7
Data fusion efficiency (%)	95	82	85	90	91
Real-time fault diagnosis accuracy rate (%)	94.2	88.5	89.6	89.3	90.7

Table 3 illustrates how GIM offers notable benefits in every area, emphasizing its fundamental technical skills in dynamic adaptation, semantic expression, and multi-source heterogeneous data processing. GIM outperforms BIM (88.2%), KG (91.7%), RF (92.5%), and LSTM (90.3%) in terms of modeling correctness, scoring 96.5%. When combining various data, GIM produces stable modeling and high restoration accuracy. With a prediction error rate of just 3.1%, it outperforms other approaches in terms of generalization and prediction robustness. In areas such as field recognition, GIM excels with a semantic consistency of 98.4% and an adaptability score of 9.2. GIM has a 95% data fusion efficiency, a 9.5-point dynamic parameter adjustment capability, and significant cross-system collaboration and self-tuning capabilities. With a 94.2% real-time fault diagnostic accuracy rate, GIM satisfies the needs for prompt recovery and early grid security alerts.

One important parameter for assessing DTPG models' real-time performance is model synchronization latency. It is essential for effective decision-making and for detecting changes in the grid state. Synchronization latency comparison studies are carried out across several models, including GIM, BIM, KG, RF, and LSTM, in order to methodically assess the synchronization performance benefits of the GIM strategy. The information gathered from several studies is described in full in Figure 6.

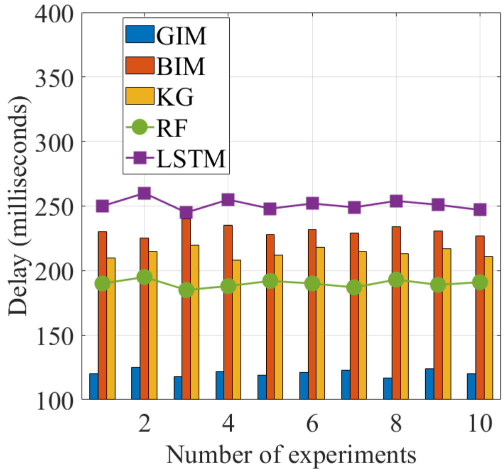

Figure 6: Comparison of synchronization delays of different DT models

The synchronization delay of GIM is consistent between 117 and 125 milliseconds, with an average of 120.9 milliseconds over 10 experiments, as illustrated in Figure 6. This indicates improved real-time responsiveness because it is substantially faster than BIM (average 231.1 milliseconds), KG (average 213.9 milliseconds), RF (average 190 milliseconds), and LSTM (average 251.1 milliseconds). This is because GIM's lightweight model design and dynamic mapping mechanism allow for quick field identification, unit unification, and semantic alignment for heterogeneous data from several sources, speeding up data synchronization. High-precision geometric modeling, which is essential to BIM, places a significant processing load on data. Because of problems with their model structure or feature extraction techniques, RF and LSTM are difficult to effectively synchronize in large-scale, real-time data streams.

5.3 Dynamic optimization performance

Improving model accuracy and early warning capabilities are core goals in power grid DT modeling and optimization. The GIM-based DTPG model utilizes a combined LSTM and RF optimization approach, integrating the advantages of time series modeling and decision tree integration to enhance the ability to capture and predict grid state changes. It is compared with the LSTM, RF, BIM, CNN-RF (Convolutional Neural Network-RF), and RNN-LSTM (Recurrent Neural Network-LSTM) algorithms, and evaluated based on the power flow calculation error (reflecting the model's ability to restore the physical state of the power grid) and the voltage over-limit warning accuracy (reflecting the practical value of predicting power grid safety risks). The experimental results are shown in Figure 7.

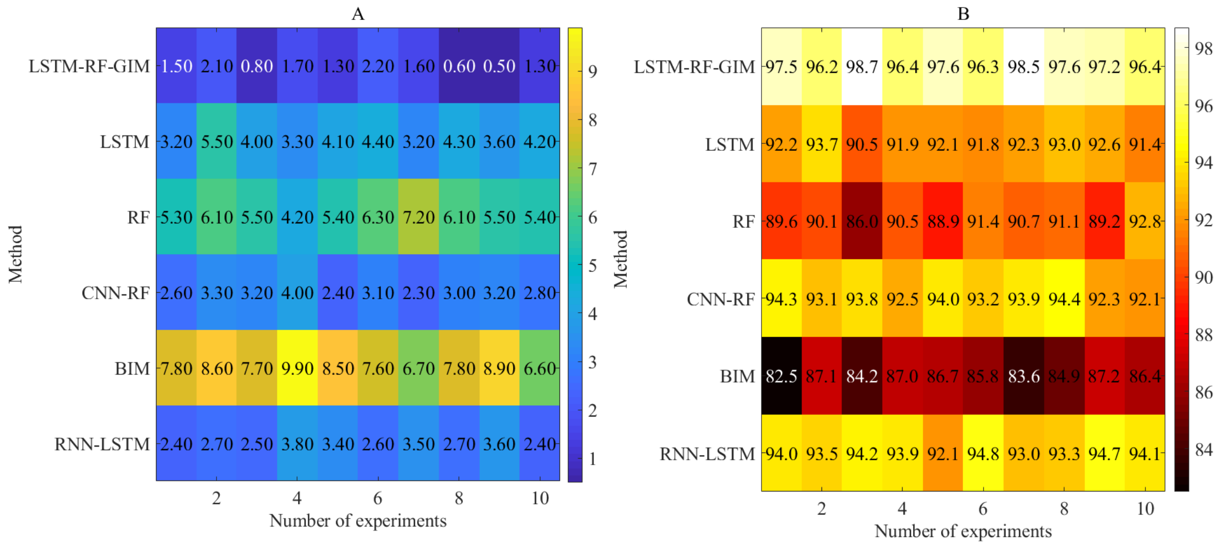

Figure 7: Performance comparison of different methods in power flow calculation and voltage warning; Figure 7A: Power Flow Calculation Error Rate (%), Figure 7B: Voltage Over-Limit Warning Accuracy (%)

As shown in Figure 7, the LSTM-RF-GIM method studied in this paper outperforms other models on the selected metrics. As shown in Figure 7A, in terms of power flow calculation error, the LSTM-RF-GIM method has a stable error rate ranging from 0.5% to 2.2%, with an average error of only 1.36%, significantly lower than LSTM (average 3.98%), RF (average 5.7%), CNN-RF (average 2.99%), BIM (average 8.01%), and RNN-LSTM (average 2.96%). As shown in Figure 7B, the LSTM-RF-GIM method achieves an accuracy rate of over 96% for voltage over-limit warnings, with a maximum of 98.7% and an average of 97.24%, significantly outperforming other methods.

5.4 Ablation experiments

We performed ablation experiments to quantitatively study the performance of Levenshtein distance, BERT semantic model, and their combination techniques in order to assess the efficacy of various similarity calculation strategies in multi-source data semantic alignment. Accuracy, F1 score, and processing time were among the evaluation measures used in the tests, which were conducted on the same test set. Table 4 displays the findings.

Table 4: Comparison results of semantic alignment ablation experiments

Method	Accuracy rate (%)	F1 score	Processing time (seconds/thousand pairs)
Use only Levenshtein	85.3	0.821	12.5
Use only BERT	91.7	0.893	68.3
Levenshtein + BERT	97.8	0.952	75.1

The ablation experiments illustrate the importance of various method characteristics and integrated tactics, as indicated in Table 4. Even while the Levenshtein distance method by itself speeds up processing, it is unsuccessful in situations when there is semantic similarity but string discrepancies because of its poor accuracy and F1 score, which make it unable to understand complicated semantics. Accuracy and F1 scores are greatly increased by the BERT semantic model, demonstrating its potent deep semantic comprehension capabilities. However, in real-time power grid applications, its high computing cost and time consumption continue to be limits. Our suggested hybrid approach, which combines Levenshtein and BERT, achieves the best possible performance-efficiency balance. It uses Levenshtein distance to quickly filter simple matching cases, reducing the computational load on BERT, while retaining the deep semantic understanding advantages of BERT to increase accuracy to 97.8%. The cognitive processes of humans are modeled by this hierarchical processing mechanism. The rationality and superiority of our suggested technique are fully validated when the combined approach provides comprehensive performance that significantly outperforms single-method solutions after making limited time trade-offs.

5.5 Significance Test

This work blends engineering usefulness with statistical significance. Significant gains in key performance indicators (KPIs) across a range of metrics were shown by the trial, suggesting statistically significant outcomes. By averaging several measures to reduce random fluctuations and doing ten repeat trials, the experimental design provided additional support for these findings. The improvements were revolutionary from an engineering standpoint: modeling time was cut to seconds, allowing for quick iterative model development. While improved field recognition accuracy reduced the need for manual intervention in data fusion, synchronized latency optimization allowed for sub-second response times. Table 5 presents the results in detail.

Table 5: Significance analysis

Performance indicators	Method of this article	Best benchmarking method	p value	Salient conclusion
Modeling time (seconds)	4.21 ± 0.12	CIM: 10.91 ± 0.19	< 0.001	Extremely significant
Synchronization delay (milliseconds)	121.5 ± 2.8	RF: 190.2 ± 5.1	< 0.001	Extremely significant
Field recognition accuracy (%)	97.82 ± 0.41	BERT: 92.10 ± 0.85	< 0.001	Extremely significant
Current calculation error rate (%)	1.36 ± 0.52	CNN-RF: 2.99 ± 0.61	< 0.01	Very significant
Modeling accuracy (%)	96.50 ± 0.75	KG: 91.70 ± 1.12	< 0.01	Very significant

All of the core performance indicators were statistically validated through hypothesis testing, as indicated in Table 5: parameter such as power flow calculation error rate-maintained p-values below 0.01 (classified as "very significant"), while three key metrics, including modeling time, achieved p-values below 0.001 (reaching the "extremely significant" level). Thorough statistical research shows that random errors are not likely to be the cause of the performance difference between our GIM approach and current benchmark methods. Our approach's superiority and cross-dimensional consistency attest to its sophisticated methodology and usefulness in engineering. The dependability and application of this research are mutually validated by the statistical significance and significant engineering advancements.

6 Discuss

The GIM-based digital twin power grid construction method proposed in this study is comprehensively superior to the existing advanced methods in terms of key performance indicators.However, in-depth analysis of the limitations of competitive methods and their own technical boundaries is very critical to objectively evaluate their contribution and applicability.The current mainstream methods have structural challenges in dealing with large-scale and complex power grids: although the general information model represented by CIM lays the foundation for the standardization of equipment description, it is difficult for static topology description and planar data models to capture the dynamic changes in the operation of the power grid. Spatial and spatial coupling relationship, real-time simulation and deduction errors are large, and the semantic conflicts of multi-source heterogeneous data cannot be automatically resolved.The modeling method based on BIM has advantages in three-dimensional geometric expression and spatial relationship characterization. It can be calculated with high cost and relies on high-precision rendering. The model synchronization delay often exceeds 200 milliseconds (the average in this experiment is 231.1 milliseconds), which is difficult to meet the low latency requirements of real-time control applications.The semantic correlation reasoning of the knowledge graph method is excellent, but the construction relies on a priori knowledge in the field, the graph update is lagging, the reasoning chain is complex, and the computational complexity is high (the delay in this experiment is 213.9 milliseconds), which limits its application in real-time decision-making scenarios.Although purely data-driven machine learning methods (such as LSTM and RF) can learn laws from historical data, the model has weak interpretability, insufficient characterization of operating rules and physical constraints, and the generalization ability drops sharply when the topology or operating mode of the power grid changes.

The GIM-based digital twin power grid construction method proposed in this study relies on a GIM-driven unified semantic framework, integrates structured domain knowledge and data-driven algorithms, and cooperatively optimizes modeling accuracy (96.5%), real-time (120.9 millisecond delay), and dynamic adaptability (parameter self-learning).However, this method has limitations.First, the core advantage relies on high-quality labeling data. The initial mapping dictionary construction of the semantic mapping layer and the fine-tuning of the BERT model require a large number of manually labeled field alignment samples, and the deployment cost of the new system is high.The second is that when the scale of the system is expanded to the national power grid, the extensibility is challenged, and the complexity of the topology of the graph database increases exponentially, affecting the efficiency of path query and island identification; Kafka event stream data throughput is large, and the network bandwidth and distributed computing architecture requirements are high.Third, the training and verification are based on the data of a single power company, and there are regional deviations. When applying across regional power grids, the semantic model and the predictive model need to be adaptively adjusted in additional fields.In the future, research will develop semi-supervised semantic alignment technology to reduce the dependence of labeling data, explore hierarchical and distributed digital twin architectures to improve extensibility, introduce a federal learning framework, use multi-regional data to improve model generalization capabilities, and promote the widespread application of technology.

7 Conclusion

This study focuses on the challenges encountered in traditional DTPG models when handling large-scale, complex power grids and multi-source, heterogeneous data fusion. It proposes a DTPG model construction and optimization method based on GIM. In terms of construction contribution, a GIM-driven full-element ontology modeling framework for power grids is developed to achieve high-precision, standardized modeling and semantically consistent management of power grid equipment. This creates a solid and semantically rich data basis for the model by successfully addressing issues like data model fragmentation and semantic redundancy. The efficiency and accuracy of heterogeneous power data fusion are also greatly improved by a five-layer multi-source data access and mapping system, which guarantees complete and up-to-date input data for the model. In terms of optimization contribution, the model can dynamically detect changes in key parameters and make real-time corrections thanks to the application of the LSTM-RF combination algorithm's self-learning optimization mechanism. This significantly improves the model's dynamic adaptability and prediction accuracy and effectively compensates for the limitations of traditional models in handling dynamic changes in the power grid. Experiments show that this approach outperforms conventional approaches in a variety of important metrics, such as the quantity of semantic ambiguities, modeling time, and field recognition accuracy. With high real-time response capabilities and a consistent synchronization delay of 117 to 125 milliseconds, the model can specifically fulfill the demands of power grid status perception and decision-making in real-time. In order to support the intelligent evolution of power systems, this study presents a technically sound methodology and techniques for the creation and optimization of DTPG models.

In the research of digital twin power grid model, there are a large number of professional terms and abbreviations involved. In order to facilitate readers to understand the key concepts in this paper, the terminology table is collated as shown in Table 6.

Table 6: Full name of abbreviations

Abbreviation	Full Form	Abbreviation	Full Form
GIM	Grid Information Model	PMU	Phasor Measurement Unit
DTPG	Digital Twin Power Grid	BIM	Building Information Modeling
LSTM	Long Short-Term Memory	KG	Knowledge Graph
RF	Random Forest	MM	Manual Modeling
BERT	Bidirectional Encoder Representations from Transformers	SMP	Static Mapping Processing
GIS	Geographic Information System	CNN	Convolutional Neural Network
SCADA	Supervisory Control and Data Acquisition	RNN	Recurrent Neural Network
JSON	JavaScript Object Notation	DFS	Depth-First Search
ETL	Extract, Transform, Load	IQR	Interquartile Range
OPC	Open Platform Communications	CPU	Central Processing Unit
RDFS	Resource Description Framework Schema	GPU	Graphics Processing Unit
OWL	Web Ontology Language	F1	F1 Score

References

[1] X. Wang, L. Wang, Y. Yu, Z. AO, and L. SUN, “Survey on characteristics, architecture and applications of digital twin power grid,” Journal of Electronics and Information Technology, vol. 44, no. 11, pp. 3721–3733, 2022. https://doi.org/10.11999/JEIT220629

[2] B. Hao and W. Yuli, “Digital power grid based on digital twin: Definition, structure and key technologies,” *Energy Reports*, vol. 8, pp. 390–397, 2022.

[3] W. Tang, X. Chen, T. Qian, G. Liu, M. Li, and L. Li, “Technologies and applications of digital twin for developing smart energy systems,” *Strategic Study of Chinese Academy of Eng*ineering, vol. 22, no. 4, pp. 74–85, 2020. https://doi.org/10.15302/J-SSCAE-2020.04.010

[4] Mchirgui, Nabil, Quadar, N., Kraiem, H., & Lakhssassi, A. "The applications and challenges of digital twin technology in smart grids: A comprehensive review." Applied Sciences 14.23 (2024):10933.https://doi.org/10.3390/app142310933.

[5] F. Dewangan, A. Y. Abdelaziz, and M. Biswal, “Load forecasting models in smart grid using smart meter information: A review,” *Energies (Basel)*, MDPI, vol. 16, no. 3, p. 1404, 2023. https://doi.org/10.3390/en16031404

[6] A. Chehri, I. Fofana, and X. Yang, “Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence,” *Sustainability*, MDPI, vol. 13, no. 6, p. 3196, 2021. https://doi.org/10.3390/su13063196

[7] D. An, F. Zhang, Q. Yang, and C. Zhang, “Data integrity attack in dynamic state estimation of smart grid: Attack model and countermeasures,” *IEEE Transactions on Automation Science and Engineering*, IEEE, vol. 19, no. 3, pp. 1631–1644, 2022. https://doi.org/10.1109/TASE.2022.3149764

[8] F. Yang, T. Wu, R. Liao, J. Jiang, T. Chen, and B. Gao, “Application and implementation method of digital twin in electric equipment,” *High Voltage Engineering*, vol. 47, no. 5, pp. 1505–1521, 2021.

[9] D. A. Mansour *et al.*, “Applications of IoT and digital twin in electrical power systems: A comprehensive survey,” *IET Generation, Transmission & Distribution*, Wiley Online Library, vol. 17, no. 20, pp. 4457–4479, 2023. https://doi.org/10.1049/gtd2.12940

[10] Z. Song *et al.*, “Digital twins for the future power system: An overview and a future perspective,” *Sustainability*, MDPI, vol. 15, no. 6, p. 5259, 2023. https://doi.org/10.3390/su15065259

[11] Z. Jiang, H. Lv, Y. Li, and Y. Guo, “A novel application architecture of digital twin in smart grid,” *J Ambient Intell Humaniz Comput*, Springer, vol. 13, no. 8, pp. 3819–3835, 2022. https://doi.org/10.1007/s12652-021-03329-z

[12] G. Juntao, Z. Jue, Z. Ningmin, Z. Hao, and J. Siliang, “Research on the application of digital twin technology in power grid engineering,” *Southern Energy Construction*, Energy China, vol. 10, no. 2, pp. 136–142, 2023. https://dx.doi.org/10.16516/j.gedi.issn2095-8676.2023.02.018

[13] Z. Jing, Q. Wang, Z. Chen, T. Cao, and K. Zhang, “Optimization of energy acquisition system in smart grid based on artificial intelligence and digital twin technology,” *Energy Informatics*,

Springer, vol. 7, no. 1, p. 121, 2024. https://doi.org/10.1186/s42162-024-00425-0

[14] G. Liu, L. Ji, and W. Liu, "Construction and Application of Digital Twin Model for Intelligent Relay Protection," *Electrical Technology*, vol. 24, no. 07, pp. 83–88, 2023.DOI：10.3969/j.issn.1673-3800.2023.07.012

[15] Y. Chen, Y. Cai, L. Tang, Y. Cao, and Y. Liu, "Local Evolution Model of Information Network for Reducing Outage Risk of Cyber-Physical System in Power Grid," *Journal of Electrical Engineering*, cjeecmp, vol. 17, no. 3, pp. 162–169, 2022. https://doi.org/10.11985/2022.03.019

[16] K. Shahid, K. Nainar, R. L. Olsen, F. Iov, M. Lyhne, and G. Morgante, "On the use of common information model for smart grid applications—A conceptual approach," *IEEE Trans Smart Grid*, IEEE, vol. 12, no. 6, pp. 5060–5072, 2021. https://doi.org/10.1109/TSG.2021.3095896

[17] H. J. Kim *et al.*, "A comprehensive review of practical issues for interoperability using the common information model in smart grids," *Energies (Basel)*, MDPI, vol. 13, no. 6, p. 1435, 2020. https://doi.org/10.3390/en13061435

[18] J. Zhang, X. Chen, S. Zhu, D. Lu, and X. Dong, "Design of 3D Visualization System for Secondary Circuit of Smart Substation Based on GIM Model," *Power System Protection and Control*, vol. 50, no. 19, pp. 179–187, 2022.

[19] J. Rong, L. Qi, H. Wu, M. Zhang, and X. Hu, "Framework for Evaluating the BIM Application Performance: A Case Study of a Grid Information Modeling System," *Sustainability*, MDPI, vol. 15, no. 15, p. 11658, 2023. https://doi.org/10.3390/su151511658

[20] Z. Shi, C. Meng, S. Zhang, S. He, and H. Zhang, "Research on construction design scheme of power transmission and transformation project based on power grid information model," *Manuf. Autom*, vol. 45, pp. 71–74, 2023.

[21] Z. Jinbo *et al.*, "Research on Transmission Line Crossing Distance Measurement Based on Grid Information Model.," *Electric Drive*, EBSCO vol. 52, no. 18, 2022. DOI: 10.19457/j.1001-2095.dqcd23281

[22] F. Hu, S. Zeng, D. Hong, S. Zhang, Y. Song, and C. Xu, "Semantic sequence labeling model of power dispatching based on deep long short term memory network," *Journal of Interconnection Networks*, World Scientific, vol. 22, no. Supp02, p. 2143018, 2022. https://doi.org/10.1142/S0219265921430180

[23] Y. Jin, Y. Zhang, W. Xue, P. Shen, Z. Jin, and T. He, "Semantic automatic annotation method based on artificial intelligence for electric power internet of things," *Internet Technology Letters*, Wiley Online Library, vol. 7, no. 2, p. e455, 2024. https://doi.org/10.1002/itl2.455

[24] A. A. P. Gomes *et al.*, "A Semantically Annotated 15-Class Ground Truth Dataset for Substation Equipment to Train Semantic Segmentation Models," *Data (Basel)*, MDPI, vol. 8, no. 7, p. 118, 2023. https://doi.org/10.3390/data8070118

[25] J. Xu, and H. Cui, "Research on Power User Information Alliance Chain Management Technology Based on Kafka Node Authentication," *Journal of Computer Applications and Software*, shcas, vol. 41, no. 11, pp. 33–38, 2024. https://dx.doi.org/10.3969/j.issn.1000-386x.2024.11.004

[26] N. Ortiz, A. A. Cardenas, and A. Wool, "Scada world: An exploration of the diversity in power grid networks," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, ACM Digital Library, vol. 8, no. 1, pp. 1–32, 2024. https://doi.org/10.1145/3639036

[27] T. Akhtar and B. B. Gupta, "Analysing smart power grid against different cyber attacks on SCADA system," *International Journal of Innovative Computing and Applications*, Inderscience, vol. 12, no. 4, pp. 195–205, 2021. https://doi.org/10.1504/IJICA.2021.116656

[28] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Intrusion detection in SCADA based power grids: Recursive feature elimination model with majority vote ensemble algorithm," IEEE Trans Netw Sci Eng, IEEE, vol. 8, no. 3, pp. 2559–2574, 2021. https://doi.org/10.1109/TNSE.2021.3099371

[29] Kumari, Namita, Sharma, A., Tran, B., Chilamkurti, N., & Alahakoon, D. "A comprehensive review of digital twin technology for grid-connected microgrid systems: State of the art, potential and challenges faced." Energies 16.14 (2023): 5525.https://doi.org/10.3390/en16145525

[30] Ruhe, Stephan, Schaefer, K., Branz, S., Nicolai, S., Bretschneider, P., & Westermann, D."Design and implementation of a hierarchical digital twin for power systems using real-time simulation." Electronics 12.12 (2023): 2747.https://doi.org/10.3390/electronics12122747

[31] A. Shukla, S. Dutta, P. K. Sadhu, and B. Dey, "An intelligent Island detection scheme to enhance grid resilience," Microsystem Technologies, Springer, vol. 30, no. 10, pp. 1363–1379, 2024. https://doi.org/10.1007/s00542-023-05602-x

[32] F. Succetti, A. Rosato, R. Araneo, G. Di Lorenzo, and M. Panella, "Challenges and perspectives of smart grid systems in Islands: a real case study," *Energies (Basel)*, MDPI, vol. 16, no. 2, p. 583, 2023. https://doi.org/10.3390/en16020583

[33] K. H. Reddy, "Modified SRFPLL harmonic support island detection in integrated electric grid," *Electrical engineering*, Springer, vol. 104, no. 3, pp. 1305–1315, 2022. https://doi.org/10.1007/s00202-021-01384-5

[34] L. Lv, Z. Wu, J. Zhang, L. Zhang, Z. Tan, and Z. Tian, "A VMD and LSTM based hybrid model of

load forecasting for power grid security," *IEEE Trans Industr Inform*, IEEE, vol. 18, no. 9, pp. 6474–6482, 2021. https://doi.org/10.1109/TII.2021.3130237

[35] D. Guha, R. Chatterjee, and B. Sikdar, "Anomaly detection using LSTM-based variational autoencoder in unsupervised data in power grid," *IEEE Syst J*, IEEE, vol. 17, no. 3, pp. 4313–4323, 2023. https://doi.org/10.1109/JSYST.2023.3266554

[36] H. Jahangir, H. Tayarani, S. S. Gougheri, M. A. Golkar, A. Ahmadian, and A. Elkamel, "Deep learning-based forecasting approach in smart grids with microclustering and bidirectional LSTM network," *IEEE Transactions on Industrial Electronics*, IEEE, vol. 68, no. 9, pp. 8298–8309, 2020. https://doi.org/10.1109/TIE.2020.3009604