Bio-Byword Scientific Publishing Pty Ltd

Journal of Architectural Research and Development, 2025, Volume 9, Issue 6
https://ojs.bbwpublisher.com/index.php/JARD
Online ISSN: 2208-3537
Print ISSN: 2208-3529

Multi-Source Heterogeneous Data Fusion Analysis Platform for Thermal Power Plants

Jianqiu Wang*, Jianting Wen, Hui Gao, Chenchen Kang

Guoteng Shanxi Hequ Power Generation Co., Ltd., Xinzhou 036500, Shanxi, China

**Author to whom correspondence should be addressed.*

Abstract: With the acceleration of intelligent transformation of energy system, the monitoring of equipment operation status and optimization of production process in thermal power plants face the challenge of multi-source heterogeneous data integration. In view of the heterogeneous characteristics of physical sensor data, including temperature, vibration and pressure that generated by boilers, steam turbines and other key equipment and real-time working condition data of SCADA system, this paper proposes a multi-source heterogeneous data fusion and analysis platform for thermal power plants based on edge computing and deep learning. By constructing a multi-level fusion architecture, the platform adopts dynamic weight allocation strategy and 5D digital twin model to realize the collaborative analysis of physical sensor data, simulation calculation results and expert knowledge. The data fusion module combines Kalman filter, wavelet transform and Bayesian estimation method to solve the problem of data time series alignment and dimension difference. Simulation results show that the data fusion accuracy can be improved to more than 98%, and the calculation delay can be controlled within 500 ms. The data analysis module integrates Dymola simulation model and AERMOD pollutant diffusion model, supports the cascade analysis of boiler combustion efficiency prediction and flue gas emission monitoring, system response time is less than 2 seconds, and data consistency verification accuracy reaches 99.5%.

Keyword: Thermal power plant; Multi-source heterogeneous data; Data fusion analysis platform; Edge computing

Online publication: Dec 10, 2025

1. Introduction

With the transformation of energy system to intelligent and efficient, thermal power plant, as the core infrastructure to ensure energy security, its equipment operation status monitoring and production process optimization has become the focus of the industry. In complex thermal cycle systems, the operation parameters, environmental data and operation and maintenance records of key equipment such as boiler, steam turbine and generator are heterogeneous from multiple sources. Traditional single-dimensional data analysis methods cannot meet the requirements of equipment life cycle management. By integrating data from physical sensors (temperature, vibration, pressure) with real-time operational data from SCADA systems, multi-source heterogeneous data fusion technology overcomes the limitations of single-source information, providing comprehensive decision support for equipment fault prediction and energy efficiency optimization [1]. For example, the fusion of infrared thermal imaging and 3d spatial data has been proven to have significant advantages in defect location and maintenance efficiency improvement in the monitoring of solar power plant equipment, and this cross-modal data integration strategy can also be applied to the state assessment of core equipment in thermal power plants.

2. Design of multi-source heterogeneous data fusion analysis platform for thermal power plants

2.1. Design of data fusion module

As the core component of the thermal power plant multi-source heterogeneous data fusion and analysis platform, the data fusion module should be designed to take into account the diversity of data collection, real-time processing and accuracy of analysis. This module establishes a multi-layered integrated architecture to enable deep interaction between physical and information spaces, ensuring effective integration of heterogeneous data from sensor networks, historical databases, and external systems. In the design process, based on the theoretical framework of data fusion, multi-dimensional data calibration and dynamic weight allocation strategy are adopted to solve the differences of different data sources in time series, format and dimension. In order to improve the data quality, the module introduces the abnormal detection algorithm based on physical constraints, and establishes the multi-parameter correlation model based on the operation mechanism of thermal power units, which can effectively eliminate the noise data and correct the measurement deviation.

In terms of implementation, the module first accesses various data sources through standardized interface protocols, including real-time sensor signals such as temperature, pressure and flow, as well as structured data from SCADA systems. To address the spatiotemporal alignment requirements of heterogeneous data, we employ time series interpolation and spatial coordinate mapping techniques to unify multi-source data under a unified spatiotemporal reference framework [2,3]. Subsequently, a 5D model based on digital twin technology establishes a mapping relationship between physical and virtual spaces. The physical model describes the unit’s operational status, while the virtual model is updated in real-time through data fusion results, enabling bidirectional information flow interaction. This model integrates physical sensor data, simulation calculation results, and expert knowledge, forming a collaborative analysis foundation for multi-source information.

2.2. Data analysis and visualization module design

As the core function unit of the multi-source heterogeneous data fusion and analysis platform for thermal power plants, the data analysis and visualization module is mainly responsible for data processing, intelligent analysis and multi-dimensional visualization. The module is built on a B/S architecture, and the hierarchical design realizes the integration of the whole process of data collection, storage, calculation and display. The technical implementation fully integrates dynamic modeling, real-time communication and interactive visualization. At the data processing level, the module design includes data preprocessing, feature extraction and standardized transformation, which can effectively deal with the format difference and noise interference of multi-source heterogeneous data in the production process of thermal power plant. The system facilitates data exchange with underlying systems like DCS and SCADA through Web service interfaces, using XML for standardized data formats. By leveraging Ajax technology, it enhances real-time responsiveness on the client side, ensuring stable and timely data collection and transmission.

The data analysis module adopts a hybrid modeling strategy, which includes traditional methods such as statistical analysis and association mining, and integrates machine learning algorithms for complex pattern recognition. For example, to meet the dynamic monitoring requirements of key parameters such as boiler combustion efficiency and steam turbine vibration characteristics, the system introduced a simulation model developed on the Dymola platform.

Through comparison and verification with measured data, the system achieved high-precision prediction of equipment operating status. For multi-source data fusion analysis, the system adopts a cascaded modular architecture. Each data processing unit performs specific functions independently: the flue gas emission monitoring module integrates the AERMOD model for pollutant diffusion simulation, while the thermal system module achieves coordinated analysis of energy and material flows through a cascaded topology [4]. This modular design not only improves the scalability of the system, but also supports flexible configuration of the analysis process according to actual needs. For example, in the scenario of joint optimization scheduling, the system can call the wind power probability prediction model and thermal power load allocation algorithm to generate multi-objective optimization scheduling scheme.

3. Data collection and processing

3.1. Data collection methods

The data collection system design of the multi-source heterogeneous data fusion analysis platform for thermal power plants follows the principles of full life cycle coverage, multi-dimensional fusion and high real-time, and builds a distributed data acquisition architecture. The data sources of the platform include various heterogeneous data sources such as unit operation parameters, environmental monitoring data, equipment status information, production management records, and external meteorological energy data. The system implements a hierarchical data collection strategy tailored to different data sources. For time-series data requiring real-time processing, it establishes bidirectional communication interfaces with power plant DCS and SCADA systems via OPC UA protocol, enabling millisecond-level acquisition of critical parameters such as temperature, pressure, and flow rate from core equipment including boilers, steam turbines, and generators; For unstructured manual inspection records and equipment maintenance logs, structured processing is performed through customized data entry interfaces, utilizing natural language processing technology to extract key information fields. For third-party data sources such as environmental monitoring systems, standardized API interfaces and HTTPS encrypted transmission protocols are employed to ensure data integrity and security during collection.

In the data acquisition process, the system sets up a three-level quality assurance mechanism. First, the redundant sensors and intelligent gateway at the hardware layer realize the preliminary verification of the data source, and the adaptive filtering algorithm is used to eliminate the high-frequency noise interference. Secondly, the QoS Level 2 service quality is implemented at the transport layer using the MQTT protocol, ensuring transmission reliability through heartbeat packet detection and data packet retransmission mechanisms. Finally, a data preprocessing module is deployed on the edge computing node to perform timestamp synchronization, dimensionality unification, and format standardization on the collected raw data. For historical data requiring protocol conversion, the system employs ETL tools to transform industrial protocols such as Modbus, Profibus into standard JSON format, while establishing a data dictionary to map physical quantities to business semantics [5].

3.2. Data preprocessing and cleaning

In the process of data standardization and normalization, appropriate transformation strategies should be selected according to the distribution characteristics of different data types. For continuous variables, Z-score normalization is employed to eliminate dimensional differences, with the formula: where μ and σ represent the sample mean and standard deviation, respectively. Discrete variables are converted into numerical features through one-hot encoding. For text-based device status description data, a structured conversion process based on natural language processing is established to extract key semantic features through regular expressions and map them to a predefined classification coding system. In terms of data quality assessment, a multi-dimensional quality index system is constructed, including key indicators such as data integrity (missing rate < 0.5%), consistency (cross-system data deviation < 3%), timeliness (delay time < 10 seconds), etc., and the credibility of data is quantified by the quality scoring model [6]. To address format heterogeneity, ETL (Extract-Transform-Load) tools are employed to standardize data formats, including XML, CSV, and database tables. A metadata catalog is established to document source information such as data origin, collection time, and sensor model.

4. Experiment and analysis

4.1. Experimental methods and steps

This study takes the typical operation scenario of thermal power plant as the background, and adopts the multi-source heterogeneous data fusion analysis platform to carry out collaborative analysis of boiler combustion system, steam turbine power generation system and environmental monitoring system. The experiment design follows the whole process framework of “data collection, preprocessing, fusion modeling, verification and evaluation”, and focuses on verifying the technical effectiveness of the platform in data space-time alignment, feature extraction and dynamic modeling.

The experimental data collection employed a hierarchical heterogeneous strategy. First, the SCADA system captured 32 types of time-series data (including boiler main steam temperature and pressure) with 1-second sampling intervals. Simultaneously, process parameters such as coal feed rate and air volume from the DCS system were sampled at 5-second intervals, while NOx emission concentration data from environmental monitoring stations were recorded every 10 minutes [7]. To ensure data integrity, a spatiotemporal reference coordinate system with multi-source data was established. All sensor devices were synchronized using GPS clocks, and data with different sampling frequencies were time-aligned through cubic spline interpolation. For unstructured data such as boiler endoscope video streams, the YOLOv5 model is employed to perform semantic segmentation of critical equipment components and extract flame morphology parameters.

4.2. Experimental results and analysis

In this experiment, a prototype system of multi-source heterogeneous data fusion analysis platform was constructed and tested in the actual operation environment of a 300MW thermal power unit, focusing on the performance of the platform in data integration efficiency, model prediction accuracy and system response performance. The experimental data included real-time sensor data from equipment such as boilers, steam turbines, and auxiliary systems, as well as historical operation records from SCADA systems, DCS control commands, and external meteorological data, with a total of about 2.5 million valid data samples collected. The cross-validation method is used to verify the performance of the fusion algorithm and compare the difference with the traditional single source data analysis method.

At the data fusion level, the platform employs ETL processes to convert data from multiple formats (including Modbus, OPC UA, and CSV) into standardized time-series sequences. The data cleansing module successfully identified and corrected 12.7% of outliers and missing data. In the feature engineering process, the sliding window method was used to extract time series features, and the fusion feature vector containing 42 dimensions was constructed by combining the statistics and frequency domain analysis. Experimental results show that after multi-source data fusion, the RMSE value of the boiler thermal efficiency prediction model decreased from 1.8% to 1.1% compared with single-source data, and the AUC value of steam turbine vibration prediction increased to 0.89, which is significantly better than the reference value of 0.76 of single sensor data [8].

5. Conclusion

In view of the urgent need for multi-source heterogeneous data fusion and analysis in thermal power plants, this study constructs an intelligent analysis platform based on edge computing and deep learning, whose core achievements are reflected in three aspects: technological system innovation, data processing capability improvement and engineering application value. By integrating data from SCADA systems, DCS control units, environmental monitoring devices, and IoT sensors, the platform has established a cross-level data fusion architecture for thermal cycles, combustion optimization, and equipment condition monitoring. This effectively resolves issues such as data silos, heterogeneous protocols, and insufficient real-time performance in traditional systems. At the data acquisition layer, a dynamic adaptation interface supporting OPC UA, Modbus TCP, MQTT and other protocols was developed to achieve millisecond-level collection and standardized conversion of key parameters of thermal power units (such as main steam temperature, boiler efficiency, pollutant emission concentration), with data integrity reaching more than 99.2%.

Disclosure statement

The authors declare no conflict of interest.

Reference

[1] Jiang J, Hong N, Zhang G, 2016, A Method for Multi-Source Heterogeneous Data Fusion and Its Application Research. Electronic Design Engineering, 24(12): 33–36.

[2] Wang Z, Liu D, Xu C, et al., 2023, Research Status and Prospects of Multi-Source Heterogeneous Data Fusion Technology in New Power Systems. China Electric Power, 56(04): 1–15.

[3] Luo G, Liu Y, Wang Z, et al., 2022, High-Precision Map Production Technology and 3D Visualization Based on Multi-Source Data. Surveying and Mapping Bulletin, 2022(11): 171–174.

[4] Bai H, Wang C, Wang Y, et al., 2025, A Method for Constructing a Big Data Warehouse for Power Grid Dispatching Error Prevention Based on Multi-Source Heterogeneous Data Fusion. Semiconductor Optoelectronics, 46(04): 750–756.

[5] Wang Z, Ge L, Wang H, 2016, Optimization Configuration Method of Feeder Automation in 10 kV Distribution Network. Journal of Power System and Automation, 28(03): 65–70.

[6] Xiong X, Xu Y, Liu X, et al., 2023, Research on Multi-source Heterogeneous Data Fusion in the Power Industry Based on Knowledge Graph Technology. China High-Tech, 2023(03): 111–112 + 132.

[7] Yang T, Wei Z, Zhou K, 2025, A Reliability Assessment Model for Multi-Source Heterogeneous Data Fusion Based on Approximate Failure Points. Mechanical Strength, 47(03): 136–142.

[8] Chen B, Chen Z, Li Z, et al., 2025, Feature Fusion of Spatial Heterogeneous Data in Power Grid Dispatching Under Multi-Layer Model. Electronic Design Engineering, 33(05): 162–165 + 171.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.