ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 22 March 2025; Accepted: 26 August 2025

https://doi.org/10.17559/TV-20250322002504
Original scientific paper

Information-Enhanced Heterogeneous Graph Convolutional Networks for Multi-Source Data Fusion in the SiC Semiconductor Industry

YiDong ZHU

Abstract: The rapid expansion of the silicon carbide (SiC) semiconductor industry has created vast and heterogeneous data streams from academic publications, patents, technical standards, and market reports. Traditional methods struggle to integrate these multi-source datasets, limiting their ability to reveal technological trends and innovation pathways. To address this challenge, we propose an Information-Enhanced Heterogeneous Graph Convolutional Network (ie-HGCN) framework for multi-source data fusion in the SiC semiconductor electronic information industry. The method combines principal component analysis and random forest algorithms for feature extraction with a DeepGCNs-Att architecture augmented by self-attention mechanisms, enabling the effective modeling of complex relationships among heterogeneous data entities. Experimental results show that ie-HGCN outperforms conventional deep learning and graph-based baselines, achieving 88.63% accuracy, 90.65% precision, 87.54% recall, and an F1-score of 85.68%. A case study on patents and publications demonstrates the framework's ability to identify emerging hotspots, such as advanced power module packaging, and to uncover novel industry-academia collaborations. These findings highlight the practical value of ie-HGCN as a robust tool for strategic R&D planning and technology forecasting in the SiC semiconductor domain.

Keywords: electronic information; identification; ie-HGCN; multi-source data; SiC semiconductor

1 INTRODUCTION

With the rapid development of technology, silicon carbide (SiC) semiconductors have been widely used in extreme environments such as high power, high frequency, and high temperature due to their excellent electronic and thermal properties [1]. In the electronic information industry, SiC devices have become a key component in fields such as power electronics, automotive electronics, and communication equipment due to their significant advantages of high efficiency, high temperature stability, and high power density. For example, in the charging stations and onboard inverters of new energy vehicles, SiC power devices significantly improve energy conversion efficiency and reduce energy losses. In the RF power amplifier of 5G base stations, SiC material helps to achieve higher power output and better signal quality [2]. However, the rapid growth of the SiC industry generates a massive and complex web of technological information. This information, characterized by its multi-source and heterogeneous nature, is critical for identifying innovation pathways. The data sources cover multiple domains such as academic literature databases (e.g., IEEE Xplore), national and international patent databases, technical standards documents, and market analysis reports. The data types include structured bibliographic data, semi-structured patent text, and unstructured technical documents. Traditional data processing methods face many serious challenges in dealing with these multi-source heterogeneous data [3, 4]. On the one hand, traditional methods are difficult to fully explore the potential correlations between data from different sources, and cannot effectively integrate the value of multi-source data (MSD), leading to a substantial decrease in the accuracy and completeness of data analysis. On the other hand, when dealing with large-scale data, traditional methods have low computational efficiency and fail to fulfill the demands for real-time processing and efficiency. For example, traditional principal component analysis (PCA) has limitations in dealing with nonlinear relationships, easily overlooking important but small variance features, which can affect the quality and reliability of data analysis [5, 6].

To address these issues, an MSD fusion method based on Information Enhanced heterogeneous graph convolutional network (ie-HGCN) technology has been proposed. The innovation of the technology lies in the construction of a scientific MSD fusion mechanism, which effectively integrates and optimizes multi-source heterogeneous data through data preprocessing, feature selection, and dimensionality reduction operations. The research also designs the ie-HGCN architecture, introduces the prediction structure of the deep graph convolutional networks (DeepGCNs) model, combined with the DeepGCNs and self attention mechanism, significantly improving the model's ability to model complex data relationships. Through these innovations, the research expects to provide an efficient and accurate solution for data fusion and processing in the SiC semiconductor electronic information industry, promoting the intelligent development of the industry.

2 LITERATURE REVIEW

Many scholars have conducted research in the field of SiC semiconductors. Ruddy F. H. et al. focused on SiC for high-temperature and high radiation nuclear detection applications. Its wide bandgap (3.27 eV) enables low-noise measurements at 700 °C and strong resistance to radiation damage. After comparing with other semiconductors, it was found that SiC detectors could still work normally after high-dose irradiation, which is suitable for advanced nuclear reactors and other fields [7]. Cheng Z. et al. were concerned about the demand for high thermal conductivity electronic materials in high-performance devices. Measurement found that the isotropic thermal conductivity of 3C SiC crystal at room temperature exceeded 500 W/m·K, and the thermal conductivity of the film was better than that of diamond film. Its high thermal conductivity was due to its high purity and high-quality structure, and it could grow in matching with the silicon lattice, providing reference for the application of next-generation power

electronic devices [8]. Shi B et al. explored the application of SiC metal-oxide-semiconductor field effect transistor (MOSFET) in electric vehicles. Compared to traditional silicon-based insulated gate bipolar transistors, SiC MOSFETs had higher operating temperatures, faster switching speeds, and higher frequencies, which could improve the performance of electric vehicles [9]. Li H. et al. studied the current imbalance problem in parallel SiC power semiconductor devices. Due to differences in circuit parameters, the current was prone to imbalance and accelerate aging when devices were connected in parallel. After analyzing the mechanism, they proposed the "imbalance degree" indicator, studied the influencing factors, summarized existing solutions, and provided insights for future technological development [10]. Kukushkin S. A. et al. reviewed the main classical growth methods, analyzed their advantages and disadvantages, and proposed a new "atomic co displacement" growth method. This method transformed the initial lattice of silicon (0.543 nm) into a cubic SiC lattice (0.435 nm), forming a nanoscale interface layer with non-standard optoelectronic properties at the SiC/silicon interface. The formation of this interface layer was due to the shrinkage of the material during the phase separation process, resulting in the structural transformation of the SiC surface area into a "semi metal". Since there were no lattice mismatch dislocations when half of the silicon atoms were replaced by carbon atoms, the SiC thin films exhibited exceptionally high crystal quality. Research showed that this SiC thin film demonstrated wide potential for application in the areas of spintronics and quantum computer components [11].

Sh L. et al. reduced customer waiting time and transportation costs through high-quality order allocation methods. To this end, a dynamic electric vehicle dispatch (DEVD) model was established, which comprehensively considered the correlation of five MSD sources: customers, vehicles, charging, stations, and services. The experimental results indicated that the proposed method outperformed the first in first out method and other ant colony optimization-based dynamic optimization algorithms in dynamic testing [12]. Li and J. studied the problem of insufficient signal penetration capability of the global positioning system (GPS) in complex environments such as urban canyons and tree canopies, resulting in positioning accuracy that cannot meet the requirements of applications such as autonomous driving. Therefore, a self localization method based on weighted cascaded compensation estimator was proposed in this study. This method integrated heterogeneous signal source (anchor) data from multiple known locations and used weighted direct localization to eliminate the non-uniformity of different signal sources, thereby accurately estimating the vehicle's position [13]. With the development of IoT devices and onboard sensors, the MSD fusion method for autonomous surface vessels has attracted attention in the intelligent edge empowered maritime IoT. A ship detection method based on lightweight YOLOX-s network was proposed, which combined transfer learning and data augmentation to achieve real-time and accurate detection of moving ships of different scales. The detection results were fused with information from the automatic recognition system to construct an augmented reality maritime navigation system, which superimposed positioning information onto video images to provide navigation risk warning for autonomous surface ships [14]. Most existing detection models only classify defects as good or bad, focusing on distinguishing between defects and intact products. Nguyen T. P. Q. et al. employed sequential clustering classification technology. Initially, they utilized traditional clustering techniques to identify defects. Subsequently, they implemented a novel approach that integrates the sine cosine algorithm with possibility fuzzy c-means for the classification and in-depth analysis of root causes. Finally, they constructed an automatic detection system using a BP neural network. In the manufacturing of pliers, the defect detection rate was 97%, the classification accuracy was 96%, and the detection cost and time were reduced [15]. Almeida V. N. D. et al. proposed a multi-objective multi-agent reinforcement learning method to address the difficulty of solving multi distributed and multi-objective problems in reality with single agent single objective reinforcement learning, as well as the problems of centralization defects and preference changes. The method allows agents to build a shared policy set in a distributed manner, and combines policy improvement and evaluation to generate behavior that adapts to any preference. The application of this method in two distinct environments demonstrated that it was capable of effectively generating agent behaviors tailored to accommodate any given target preference [16].

To sum up, numerous experts have undertaken extensive research into the utilization of SiC semiconductors in high-temperature and high-radiation settings, as well as their potential in the electronic information industry, especially in advanced nuclear reactors, electric vehicle power semiconductor devices, and other fields. However, most studies have not fully considered the complex correlations between different data sources when dealing with MSD fusion, resulting in limited performance in high-dimensional data analysis and application scenarios. Therefore, an MSD fusion method based on ie-HGCN technology is proposed in this study. By modeling the relationships between different types of nodes and edges, multi-dimensional information flows can be efficiently transmitted within the graph convolutional layer to enhance the ability to express complex data relationships. The primary methodological contribution of this work lies not in the invention of a new algorithmic layer, but in the synergistic integration of established techniques into a cohesive, domain-informed architecture tailored specifically for the SiC semiconductor industry. The proposed framework is purpose-built to address the unique challenges of SiC data, where traditional methods often fail. It uniquely combines a two-stage feature engineering process to handle high-dimensional sensor data while preserving critical low-variance parameters, with a deep graph network that explicitly models the physical and logical relationships of the SiC value chain. This integrated pipeline represents a novel approach that moves beyond a generic application of AI to provide an effective and targeted solution for data fusion and analysis in this critical domain.

3 RESEARCH METHODOLOGY

3.1 MSD Fusion Method for SiC Semiconductor Electronic Information Industry

SiC materials have excellent electronic and thermal properties, but to improve their semiconductor device performance and accelerate research and development, it is necessary to integrate heterogeneous data to extract valuable information. MSD fusion technology combines data from different sources and types to form large-scale databases [17, 18]. The data collection and preprocessing are shown in Fig. 1.

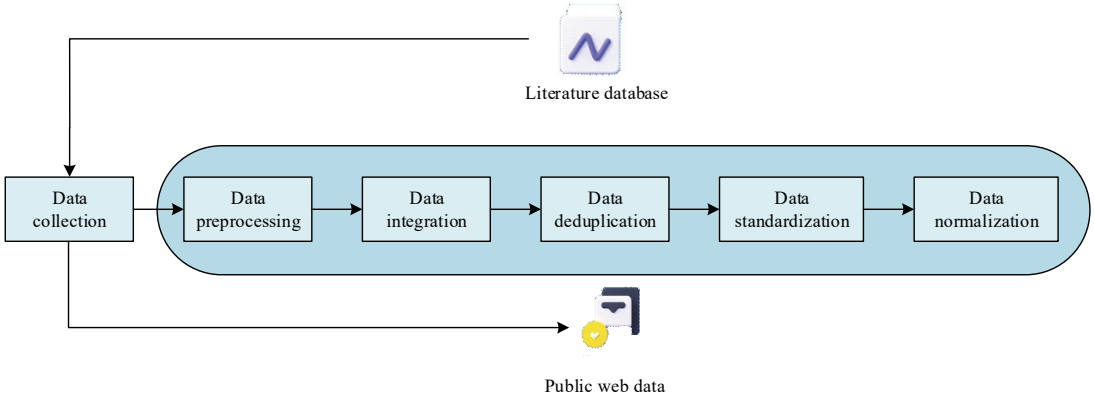


Figure 1 Data collection and preprocessing

Fig. 1 indicates the data collection and preprocessing process. Data includes material properties, device performance, and application cases, which are the foundation for establishing an industrial data system. After data collection, data preprocessing is carried out, including cleaning, integration, deduplication, standardization, and normalization. Missing and outlier values are handled during cleaning, as data accuracy is crucial for analysis and decision-making. The process establishes a unified field mapping for data from different sources during integration to facilitate effective integration. In the deduplication stage, data uniqueness is ensured through a unique identifier. The field format is standardized during the standardization stage for easy analysis. To analyze the SiC semiconductor technological landscape, this study constructs a multi-source heterogeneous network model, as shown in Fig. 2.

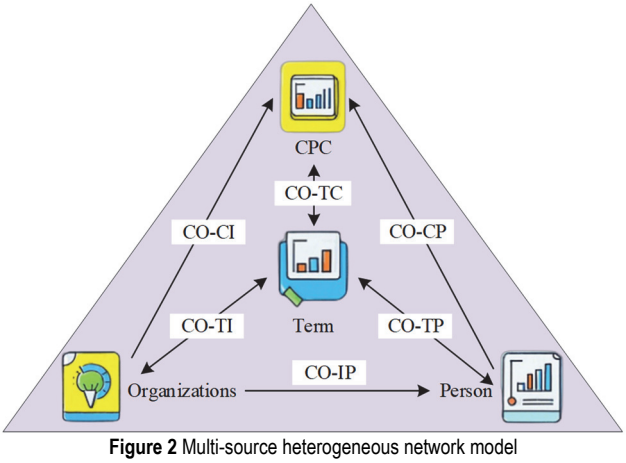


Figure 2 Multi-source heterogeneous network model

As shown in Fig. 2, this study analyzes the SiC semiconductor field using a heterogeneous network model to map the innovation ecosystem and relationships among researchers, institutions, and technology fields, categorized by CPC codes. The model includes edges for intellectual property collaboration (CO-IP), citation linkage (CO-CI), and technology-personnel links (TP), facilitating the identification of key collaborations and central technology areas in the industry's development. To ensure that features from different sources are comparable, the data are first standardized. Eq. (1) describes this process, where each data point is adjusted so that the feature set has a mean of 0 and a standard deviation of 1. This prevents features with larger scales from dominating the analysis.

$$X_{\text{standard}} = \frac{X - \mu}{\sigma} \tag{1}$$

In eq. (1), X represents the raw data. μ is the mean. σ is the standard deviation. After standardization, the mean of the data is 0 and the standard deviation is 1. After standardization, the covariance matrix of the feature matrix is calculated, as presented in Eq. (2).

$$C = \frac{1}{n-1}(X_{\text{standard}}^{T} X_{\text{standard}}) \tag{2}$$

In Eq. (2), n is the sample number. X_{standard} is the standardized feature matrix. The covariance matrix undergoes eigenvalue decomposition to yield eigenvalues and their respective eigenvectors. Based on the magnitudes of these eigenvalues, the eigenvectors corresponding to the top k eigenvalues are typically chosen. Generally, eigenvectors are selected such that their cumulative variance contribution rate falls within the range of 85% to 95%, as illustrated in Eq. (3).

$$CV = \sum_{i=1}^{k} \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{3}$$

In Eq. (3), λ_i is the i-th eigenvalue. p is the total number of features. Finally, the original standardized data are multiplied by the selected feature vector to obtain the principal components, as shown in Eq. (4).

$$Z = X_{\text{standard}} W \tag{4}$$

In Eq. (4), W represents the selected feature vector combination. Z is the main component feature matrix. PCA is effective in feature selection and dimensionality reduction in the SiC semiconductor electronic information industry. However, PCA has limitations in handling nonlinear relationships and neglects important but low variance features. To address these issues, the research employs the Random Forest (RF) algorithm [19, 20]. In essence, Eq. (2) through Eq. (4) describe the PCA process. Eq. (2) calculates the covariance matrix, which measures how different features vary together. By decomposing this matrix, a set of "principal components" is obtained. These are new, composite features that capture the most significant variance in the data. Eq. (3) provides a criterion for selecting the most important components, ensuring that most of the original information (e.g., 95%) is retained. Finally, Eq. (4) shows how the original data is projected onto these new principal components, resulting in a dataset with lower dimensions that is easier to analyze yet informationally rich. SiC manufacturing data often involves high-dimensional, highly-correlated sensor readings from processes like epitaxial growth. PCA is first applied to reduce noise and extract macro-level features, transforming raw sensor data into a smaller set of orthogonal components representing key process stages. However, PCA may miss low-variance yet critical features, such as impurity levels or substrate defects, which significantly impact device performance. Therefore, a second stage using RF incorporates both the original low-dimensional parameters and PCA-derived features. RF's built-in feature importance mechanism evaluates all inputs equally, effectively identifying the most predictive variables. This hybrid approach balances dimensionality reduction with the retention of essential low-variance features. The flowchart of the RF algorithm is shown in Fig. 3.

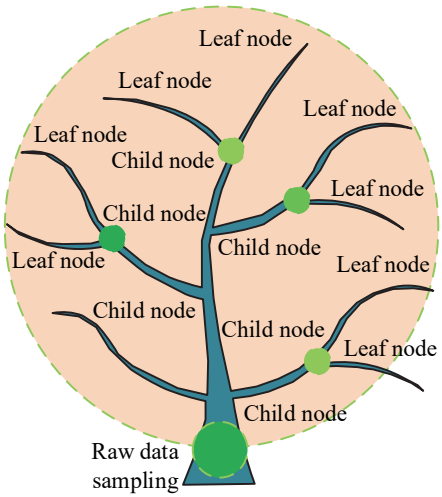


Figure 3 Flowchart of RFalgorithm

As shown in Fig. 3, the RF algorithm randomly extracts multiple subsets from the original training dataset during the data sampling process, and generates different training sets through sampling methods. A decision tree is built for each subset of samples. When splitting at each node, instead of using all features, the algorithm randomly selects some features for optimal splitting to reduce the correlation between features. Then, the principal components extracted using PCA are shown in Eq. (5).

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} T_i(Z) \tag{5}$$

In Eq. (5), $T_i(Z)$ represents the predicted value of the i-th decision tree. N is the number of trees. $\hat{Y}$ is the final prediction result of the RF for the sample. Eq. (5) illustrates the core principle of the RF algorithm. Rather than relying on a single model, it aggregates the predictions from a large ensemble of individual decision trees (k). This "wisdom of the crowd" approach makes the final prediction more robust and less prone to overfitting. After training, the feature importance evaluation mechanism embedded in the RF is used to assess the contribution of the original features and PCA principal components to the model's predictive performance, as shown in Eq. (6).

$$I(f_j) = \frac{1}{N}\sum_{i=1}^{N}\left(mse_d - mse_p\right)_i \tag{6}$$

In Eq. (6), mse_d is the mean square error under default conditions. mse_p is the mean square error recalculated after the feature f_j is randomly shuffled. Eq. (6) explains how the RF algorithm evaluates the importance of a given feature (i). The logic is intuitive: the importance is measured by calculating how much the model's prediction error increases when the values of that specific feature are randomly shuffled. A large increase in error implies that the model relies heavily on that feature, making it highly important.

3.1.1 Design of ie-HGCN Technology Based on MSD Fusion

After discussing the key role of MSD fusion in improving device performance and R&D efficiency in the SiC semiconductor electronic information industry, the research further explored the design of ie-HGCN technology. This technology enhances the model's expressive ability by fusing MSD, thereby optimizing decision output. The basic network architecture of ie-HGCN includes input layer, heterogeneous graph construction module, graph convolutional layer, information enhancement layer, and output layer [21, 22]. A high-level overview of the proposed ie-HGCN model architecture is presented in Fig. 4.

Fig. 4 illustrates the overall data processing pipeline. The process begins with MSD entering the Input Layer. This data is then used by the Heterogeneous Graph Construction Module to build the graph structure that represents the ecosystem. The core of the model, the graph convolutional layer and the information enhancement layer, then process this graph to learn node representations. Finally, the output layer uses these learned representations to make predictions, such as identifying key technology fields or influential institutions.

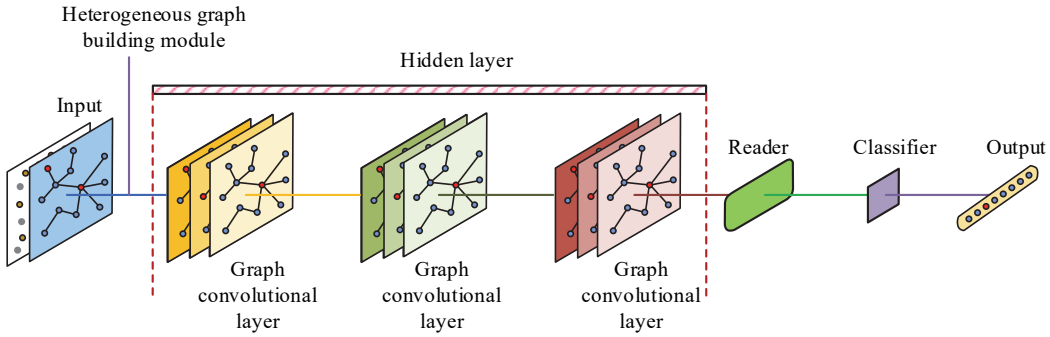


Figure 4 High-level architecture of the ie-HGCN model

As shown in Fig. 4, the architecture includes an input layer, a heterogeneous graph construction module, a graph convolutional layer, an information enhancement layer, and an output layer. The input layer receives multi-dimensional features from multiple sources, such as production equipment, market feedback, and test results. The heterogeneous graph construction module constructs a graph based on the relationship between nodes and edges. Node types and characteristics determine the information transmission method, while edge types define the relationships between nodes. Graph convolutional layers update node features through feature propagation, focus on neighborhood features, and introduce cross neighborhood information flow. The information enhancement layer integrates neighborhood information and dynamically learns feature weights to optimize the output. The output layer maps features into decision results, such as product quality ratings or market supply and demand forecasts. This architecture reflects the physical and informational value chain of the SiC industry. The heterogeneous graph acts as a digital twin, with nodes like "material batches," "CVD equipment," "testing protocols," and "research papers," connected by domain-specific relationships (e.g., "processed_by", "governed_by", "theoretical_guidance_for "). This structure enables the model to learn directly from the inherent logic of SiC manufacturing.

The use of DeepGCNs-Att addresses two key challenges in SiC data. First, long-range dependencies - defects in final devices may originate from early-stage anomalies, requiring DeepGCNs to propagate information across many graph steps. Second, varying data relevance - while final test data is critical for performance prediction, fault diagnosis may hinge on specific tool data at a precise time. The attention mechanism enables dynamic weighting of different node types based on context, enhancing prediction accuracy. In the design of ie-HGCN technology based on MSD fusion, the prediction structure of DeepGCNs-Att model is introduced to enhance the expression ability of the model. This structure combines DeepGCN and Attention Mechanism to enhance the model's ability to model complex data relationships, as shown in Fig. 5.

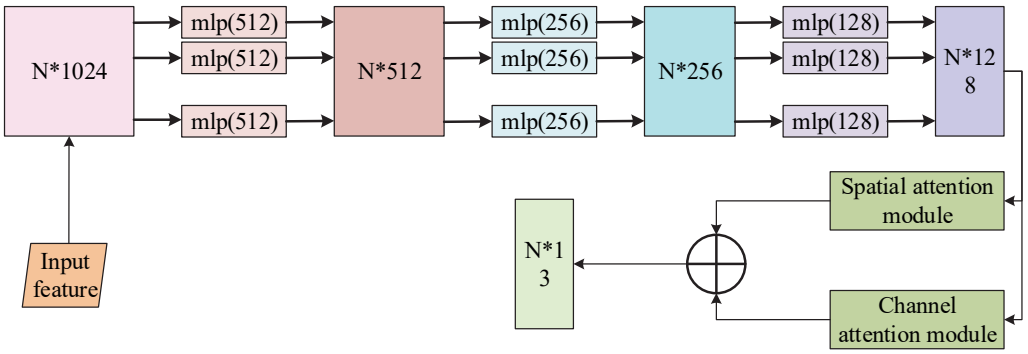


Figure 5 DeepGCNs-Att model prediction module structure

The model input in Fig. 5 is the feature data of N1024. These data are passed through three MLP (512) layers and then merged into the feature vector of N512. Next, these vectors are passed through two MLP (256) layers to output N256 features. Afterwards, the data are processed through two MLP (128) layers to obtain the features of N128 [23, 24]. Deep GCN is a model that stacks multiple graph convolutional layers and extracts local neighborhood information layer by layer to obtain high-level feature representations, as shown in Eq. (7).

$$H^{(l+1)} = \sigma\left(D^{-1/2} A D^{-1/2} H^{(l)} W^{(l)}\right) \tag{7}$$

In Eq. (7), $H^{(l)}$ represents the node feature matrix of layer l. A is the adjacency matrix. $W^{(l)}$ is the learnable weight matrix for layer l. σ is the activation function. The depth of the DeepGCN structure gives it stronger expressive power, but when the layers are too deep, the features may become overly smooth, making it difficult to transmit information. Therefore, on this basis, self attention mechanism is introduced to alleviate this problem. For each node, its attention weight compared to other nodes is calculated. Assuming that the feature of node v_i is h_i, the attention weight calculation is shown in Eq. (8) [25].

$$e_{ij} = \text{LeakyReLU}\left(a^T \left[Wh_i \,\|\, Wh_j\right]\right) \tag{8}$$

In Eq. (8), a is the learned weight vector. Then it is normalized using the softmax function, as shown in q. (9).

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(e_{ik}\right)} \tag{9}$$

In Eq. (9), $\mathcal{N}(i)$ denotes the set of neighboring nodes of node i. Then, based on attention weights, the features of neighboring nodes are weighted and summed up, as shown in Eq. (10).

$$h_i^{'} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j\right) \tag{10}$$

In Eq. (10), $h_i^{'}$ is the feature representation of node v_i after being processed by the self attention mechanism. After completing the design of the prediction module based on DeepGCNs-Att, this study further integrated all key components to form a complete ie HGCN model based on MSD fusion. In the node feature update of each layer, new node representations are obtained by aggregating the features of neighboring nodes. The self-attention mechanism, described in Eq. (8) through Eq. (10), enhances the graph convolution process. The goal is to allow a node to weigh the importance of its neighbors differently. Eq. (8) first calculates a raw attention score between a node i and its neighbor j. Eq. (9) then normalizes these scores using the softmax function, effectively creating a probability distribution of importance over the neighbors. Finally, Eq. (10) calculates the node's new feature representation by taking a weighted average of its neighbors' features, where the weights are the attention scores just calculated [26]. Assuming the eigenvector of node v_i is $H_j^{(l)}$, feature aggregation is achieved by weighting and summing the features of neighboring nodes v_j, as presented in Eq. (11).

$$H_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} H_j^{(l)} W^{(l)}\right) \tag{11}$$

In Eq. (11), σ is the activation function. c_{ij} is the edge weight between node i and node j. $W^{(l)}$ is the weight matrix. In order to enhance the flexibility and expressive power of the model in processing complex data, ie-HGCN introduces multi-scale aggregation, which enables the model to understand features in multiple contexts through different levels of aggregation mechanisms. The ie-HGCN model calculation model is shown in Fig. 6.

Fig. 6 provides a more detailed look at the calculation process for a single 'Personnel' node within a graph layer. The process involves two main steps. First, in the Object Aggregation stage, the model gathers information from different types of neighboring nodes (e.g., an "Institution" node and a "Technology" node). Second, in the Type Aggregation stage, this aggregated information from different sources is combined with the personnel node's own information. Crucially, throughout this process, learnable weight matrices are employed to empower the model to make intelligent decisions regarding the significance to be attributed to each individual piece of information. This two-step aggregation mechanism enables the creation of a rich, context-aware representation for each node in the graph.

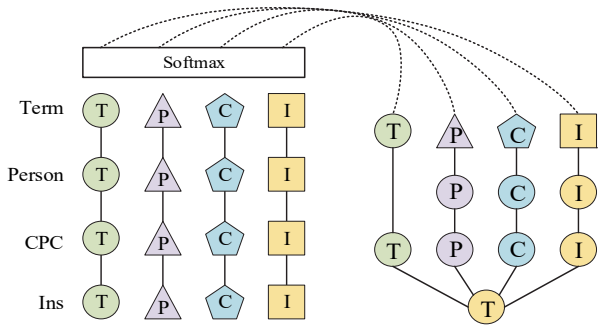


Figure 6 Calculation model of ie-HGCN model

4 RESULTS AND DISCUSSION

4.1 Experimental Setup and Dataset

To validate the performance of the proposed ie-HGCN model, a representative dataset was constructed. The empirical data presented in this study were synthesized to reflect the typical performance characteristics and statistical distributions of real-world SiC devices. The generation process was as follows:

First, an extensive survey of publicly available datasheets from leading semiconductor manufacturers and relevant academic literature was conducted to establish realistic value ranges, means, and standard deviations for key performance metrics (e.g., power density, thermal stability, efficiency, failure rate, breakdown voltage).

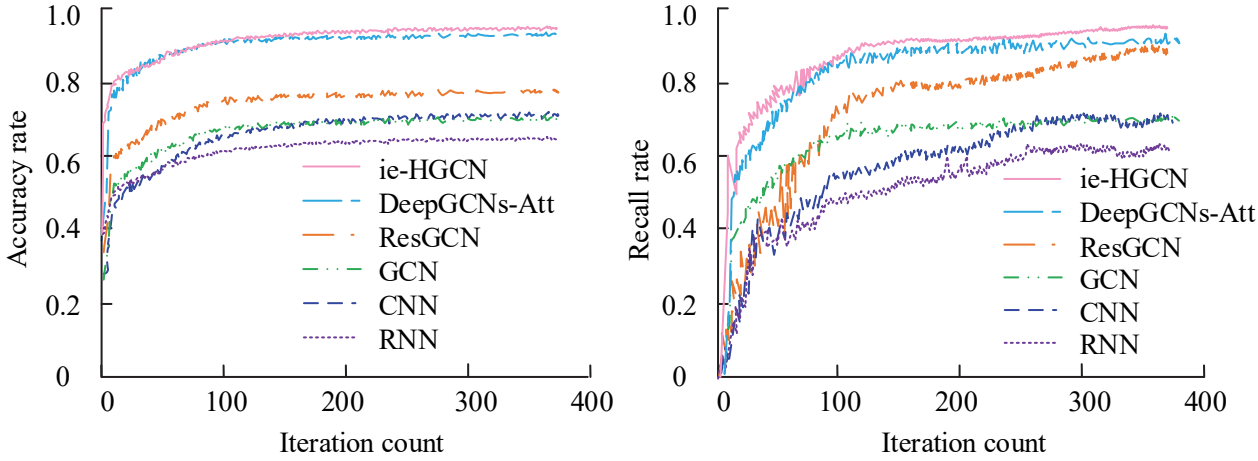
Second, based on these established statistical parameters, a larger validation dataset of several hundred data points was synthetically generated. This process ensured that the inter-metric correlations within the generated data (e.g., the typical trade-off between breakdown voltage and on-state resistance) were consistent with known physical principles of SiC devices.

Finally, the data presented in Tabs. 1, 2, and 3 were representative samples drawn from this larger synthetic validation set. This approach allows for a controlled and repeatable evaluation of the model's data fusion and analysis capabilities on a dataset that mirrors the complexity and heterogeneity of real-world SiC industry data, while avoiding the disclosure of proprietary, sensitive experimental data. All subsequent performance evaluations of the models are conducted on this validation dataset.

4.2 Performance Testing of ie-HGCN Technology Based on MSD Fusion

In the SiC semiconductor electronic information industry, robust and reliable data analysis is crucial. To rigorously evaluate the proposed ie-HGCN algorithm, its performance was benchmarked against several other models. The evaluation was conducted using a 5-fold cross-validation methodology on the validation dataset to ensure the stability and generalizability of the results. The

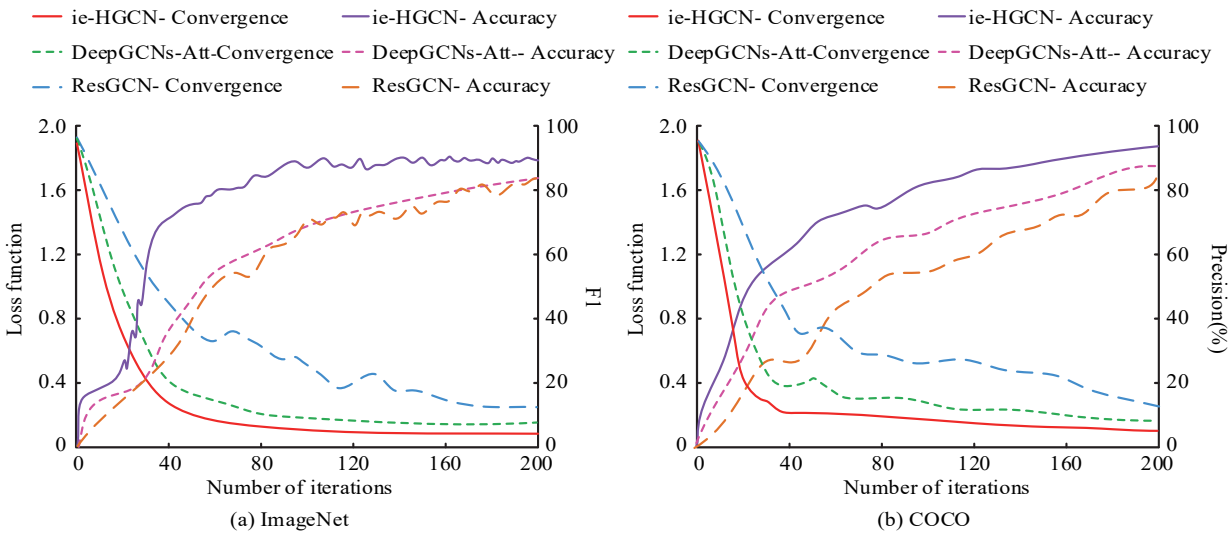
performance curves depicted in Figs. 7, 8, and 9 illustrate the convergence behavior from a single, representative training run, while the final reported metrics are the average values computed across all five folds. In the experiment, ie-HGCN was compared with DeepGCNs-Att, Residual Graph Convolutional Network (ResGCN), Graph Convolutional Network (GCN), CNN, and Recurrent Neural Network (RNN). All algorithms were trained and tested on the same dataset to ensure fairness of the results. The performance comparison of different algorithms in accuracy and recall is shown in Fig. 7.



(a) Accuracy rate (b) Recall rate

Figure 7 Comparison of accuracy and recall performance of different algorithms

Fig. 7 shows the changes in accuracy and recall of different algorithms during the iteration process. Fig. 7a shows that the accuracy of each algorithm increased with the number of iterations. Among them, the ie-HGCN algorithm had the fastest accuracy growth, reaching 88.63%, and after about 100 iterations, its accuracy remained at a high level, significantly better than other algorithms. Fig. 7b shows the variation of recall rates of each algorithm with the number of iterations. The recall rate of ie-HGCN algorithm also performed well, reaching 87.54%, and its growth rate and final recall rate were higher than other algorithms. The convergence effect and detection accuracy of the model are shown in Fig. 8.

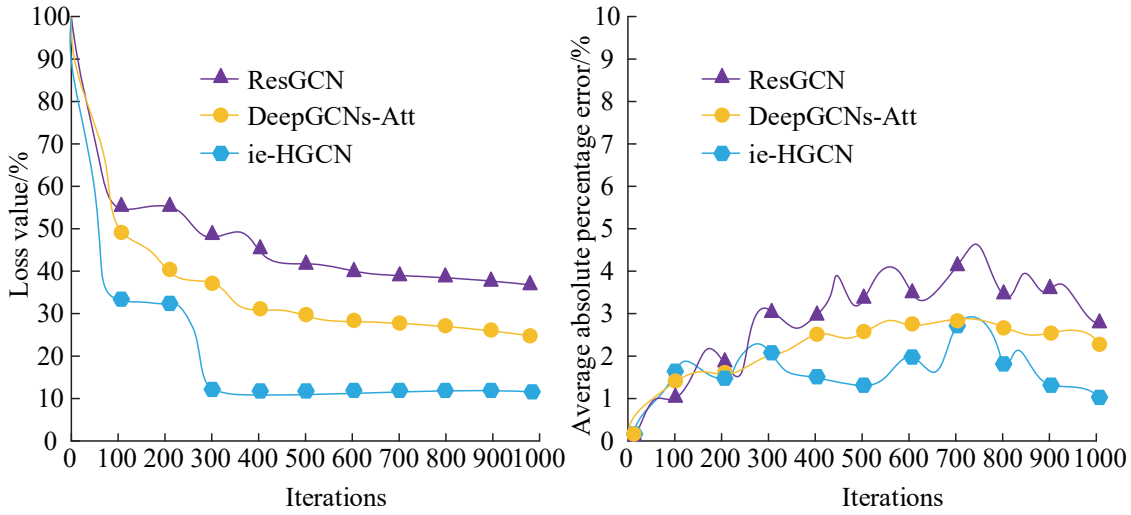


(a) ImageNet (b) COCO

Figure 8 Convergence effect and detection accuracy of the model

The convergence behavior shown in Fig. 8 provided a visual representation of the model's learning efficiency. In Figure 8(a), the loss curve of the ie-HGCN model demonstrated a rapid decrease, while its F1 value curve rose swiftly to the highest level among the compared models. Similarly, Fig. 8b shows the model achieving a high accuracy level quickly and maintaining its superiority. The performance comparison of different algorithms in training loss and average absolute percentage error is shown in Fig. 9.

Fig. 9 offers further insight into the model's stability and error rates during training. The training loss for the ie-HGCN model, depicted in Fig. 9a, was consistently lower and converged faster than the ResGCN and DeepGCNs-Att models. Moreover, Fig. 9b indicates that the ie-HGCN model maintained the lowest average absolute percentage error throughout the iteration process, ultimately stabilizing at a significantly better level than the other models.



(a) Three models for training loss functions (b) Average percentage error of three methods

Figure 9 Comparison of performance of different algorithms in training loss and average absolute percentage error

4.3 Empirical Application: Identifying Technological Hotspots

To demonstrate the model's practical utility for its defined objective, a case study was conducted. A heterogeneous graph was constructed using a dataset of SiC-related patent records and academic publications from the last decade. The ie-HGCN model was then tasked with analyzing this network to identify emerging technological hotspots, defined as technology fields showing a rapid increase in research activity and influence.

The model's analysis yielded several actionable insights. For example, it successfully identified "advanced power module packaging for high-temperature operation" as a key emerging hotspot. This was determined by the model detecting a significant increase in patent filings in this domain, coupled with a growing number of citations from foundational research on SiC substrate growth. Furthermore, the model highlighted a strong, previously non-obvious link between institutions specializing in materials science and companies filing patents in automotive power systems, pinpointing a key area of industry-academia collaboration. This case study confirmed that the ie-HGCN framework could effectively translate complex, multi-source textual data into a strategic map of the innovation landscape, directly fulfilling the core research objective. The final performance metrics, averaged across the 5-fold cross-validation, are summarized in Tab. 1.

Table 1 Overall performance comparison of different models (mean±SD %)

Model	Accuracy	Precision	Recall	F1 value
ie-HGCN	88.63 ± 0.45	90.65 ± 0.38	87.54 ± 0.51	85.68 ± 0.62
DeepGCNs-Att	80.15 ± 0.88	82.34 ± 0.91	79.55 ± 1.03	80.92 ± 0.95
ResGCN	78.50 ± 1.12	80.21 ± 1.05	77.89 ± 1.25	79.03 ± 1.18
GCN	74.22 ± 1.35	76.88 ± 1.29	73.10 ± 1.44	74.94 ± 1.39
CNN	70.56 ± 1.58	72.13 ± 1.66	69.83 ± 1.71	70.96 ± 1.62
RNN	68.91 ± 1.83	70.49 ± 1.75	68.12 ± 1.90	69.28 ± 1.81

The results in Tab. 1 indicated that the ie-HGCN model consistently outperformed the other methods across all evaluation criteria. To determine if the observed performance improvements of the ie-HGCN model were statistically significant, a Wilcoxon signed-rank test was conducted. The test compared the cross-validation results of ie-HGCN against each of the other models. The resulting *p*-values are presented in Tab. 2.

Table 2 Statistical significance test results (p-values) for ie-HGCN vs. other models

Comparison Pair	Accuracy	Precision	Recall	F1 value
ie-HGCN vs. DeepGCNs-Att	0.009	0.007	0.011	0.015
ie-HGCN vs. ResGCN	0.008	0.006	0.009	0.012
ie-HGCN vs. GCN	< 0.001	< 0.001	< 0.001	< 0.001
ie-HGCN vs. CNN	< 0.001	< 0.001	< 0.001	< 0.001
ie-HGCN vs. RNN	< 0.001	< 0.001	< 0.001	< 0.001

The results in Tab. 2 consistently showed *p*-values well below the 0.05 significance threshold across all metrics and all comparison pairs. This provided strong statistical evidence to reject the null hypothesis and conclude that the superior performance of the ie-HGCN model was not due to random chance.

5 DISCUSSION

The results presented in Section 4 demonstrated the strong performance of the ie-HGCN model, but their true significance lied in their interpretation and practical implications. The model's high precision (90.65%) was particularly valuable for strategic planning, as it minimized

the risk of allocating R&D resources to false trends. The strong recall (87.54%) ensured that companies were unlikely to miss genuine emerging technologies, which was critical for maintaining a competitive edge. Overall, the balanced and statistically significant performance confirmed that the model was a reliable and robust tool for technology scouting.

Furthermore, the empirical case study, which identified "advanced power module packaging" as a hotspot, illustrated the model's direct applicability. From a strategic management perspective, such insights were directly actionable. For instance, a Chief Technology Officer could use these findings to justify redirecting research funding, while a venture capital analyst could identify promising startups for investment. The model, therefore, acts as a bridge between raw data and strategic execution.

Despite these positive outcomes, the study has certain limitations. The primary constraint is the model's computational intensity when applied to extremely large-scale datasets, which could pose a challenge for organizations with limited computing infrastructure. This points to a need for future work on algorithmic optimization.

6 CONCLUSION

This study introduced an Information-Enhanced Heterogeneous Graph Convolutional Network (ie-HGCN) framework for multi-source data fusion in the SiC semiconductor electronic information industry. By integrating PCA and random forest feature engineering with a DeepGCNs-Att architecture enhanced by self-attention, the model effectively captured complex relationships within heterogeneous datasets. Experimental evaluation demonstrated that ie-HGCN significantly outperformed existing deep learning and graph-based methods, achieving 88.63% accuracy, 90.65% precision, 87.54% recall, and an F1-score of 85.68%. The empirical case study further validated the framework's practical utility, successfully identifying emerging technological hotspots and revealing new patterns of collaboration across institutions and industry sectors.

The contributions of this work are both methodological and practical: it establishes a domain-specific graph-based pipeline for integrating large-scale heterogeneous data, and it provides a validated decision-support tool for strategic R&D planning in the SiC industry. Nevertheless, the model's computational demands present a challenge for extremely large-scale datasets, suggesting the need for future work on optimization strategies such as graph sampling or distributed training. Further research could also extend the framework to incorporate financial, market, and temporal data, thereby enabling continuous monitoring of technology lifecycles and more comprehensive forecasting.

In sum, the proposed ie-HGCN approach offers a robust and scalable foundation for advancing intelligent data fusion in the SiC semiconductor sector, supporting both technological innovation and strategic decision-making.

7 REFERENCES

[1] Wang, Y., Hung, J. C., Huan, C., Hussain, S., Yen, N., & Jin, Q. (2024). Design of TAM-based Framework for Credibility and Trend Analysis in Sharing Economy: Behavioral Intention and User Experience on Airbnb as an Instance. *Computer Science and Information Systems*, *21*(2), 547-568. https://doi.org/10.2298/CSIS230323010W

[2] Duan, H. W., Zhang, L. P., Gan, B., Chang, X., Wang, X. F., & Li, K. H. (2023). A System Dynamics-Based Simulation Model for Cross-Border Logistics Risk Transmission. *International Journal of Simulation Modelling*, *22*(3), 485-496. https://doi.org/10.2507/IJSIMM22-3-CO11

[3] Wang, S. L. & Zhang, X. (2024). Optimization Strategies and Simulation of Integrated Management in Supply Chains. *International Journal of Simulation Modelling*, *23*(3), 543-554. https://doi.org/10.2507/IJSIMM23-3-CO15

[4] Qiu, Y. (2024). Evaluation of the Innovation Capability of Import Borderless E-commerce Platform Based on Factor Analysis and TOPSIS Method. *Tehnicki vjesnik-Technical Gazette*, *31*(3), 1011-1020. https://doi.org/10.17559/TV-20231010001012

[5] Wu, L. J., Chen, Z. G., Chen, C. H., et al. (2022). Real environment-aware multisource data-associated cold chain logistics scheduling: A multiple population-based multiobjective ant colony system approach. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 23613-23627. https://doi.org/10.1109/TITS.2022.3203629

[6] Zhao, W. H., Zhang, J. Q., Zhang, H. X., Wang, J., & Jia, R. (2022). 2D boron carbide, carbon nitride, and silicon carbide: A theoretical prediction. *ACS Applied Electronic Materials*, *4*(10), 4903-4911. https://doi.org/10.1021/acsaelm.2c00825

[7] Ruddy, F. H., Ottaviani, L., Lyoussi, A., Destouches, C., Palais, O., & Reynard-Carette, C. (2022). Silicon carbide neutron detectors for harsh nuclear environments: A review of the state of the art. *IEEE Transactions on Nuclear Science*, *69*(4), 792-803. https://doi.org/10.1109/TNS.2022.3144125

[8] Cheng, Z., Liang, J., Kawamura, K., Zhou, H., Asamura, H., Uratani, H., & Cahill, D. G. (2022). High thermal conductivity in wafer-scale cubic silicon carbide crystals. *Nature communications*, *13*(1), 7201-7210. https://doi.org/10.1038/s41467-022-34943-w

[9] Shi, B., Ramones, A. I., Liu, Y., Wang, H., Li, Y., Pischinger, S., & Andert, J. (2023). A review of silicon carbide MOSFETs in electrified vehicles: Application, challenges, and future development. *IET Power Electronics*, *16*(12), 2103-2120. https://doi.org/10.1049/pel2.12524

[10] Li, H., Zhao, S., Wang, X., Ding, L., & Mantooth, H. A. (2023). Parallel connection of silicon carbide MOSFETs - Challenges, mechanism, and solutions. *IEEE Transactions on Power Electronics*, *38*(8), 9731-9749. https://doi.org/10.1109/TPEL.2023.3278270

[11] Kukushkin, S. A. & Osipov, A. V. (2022). Epitaxial silicon carbide on silicon. Method of coordinated substitution of atoms (a review). *Russian Journal of General Chemistry*, *92*(4), 584-610. https://doi.org/10.1134/S1070363222040028

[12] Shi, L., Zhan, Z. H., Liang, D., & Zhang, J. (2022). Memory-based ant colony system approach for multi-source data associated dynamic electric vehicle dispatch optimization. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 17491-17505. https://doi.org/10.1109/TITS.2022.3150471

[13] Li, J., Li, P., Li, P., Tang, L., Zhang, X., & Wu, Q. (2022). Self-position awareness based on cascade direct localization over multiple source data. *IEEE Transactions on Intelligent Transportation Systems*, *25*(1), 796-804. https://doi.org/10.1109/TITS.2022.3170465

[14] Liu, R. W., Guo, Y., Nie, J., Hu, Q., Xiong, Z., Yu, H., & Guizani, M. (2022). Intelligent edge-enabled efficient multi-source data fusion for autonomous surface vehicles in

maritime internet of things. *IEEE Transactions on Green Communications and Networking*, *63*(5), 1574-1587. https://doi.org/10.1109/TGCN.2022.3158004
[15] Rui, G. & Li, M. (2024). Utilizing Internet Big Data and Machine Learning for Product Demand Forecasting and Analysis of its Economic Benefits. *Tehnicki vjesnik-Technical Gazette*, *31*(4), 1385-1394. https://doi.org/10.17559/TV-20240318001408
[16] Ionescu, S. A. & Diaconita, V. (2023). Transforming Financial Decision-Making: The Interplay of AI, Cloud Computing and Advanced Data Management Technologies. *International Journal of Computers Communications & Control*, *18*(6), 5735. https://doi.org/10.15837/ijccc.2023.6.5735
[17] Garcia Gastelum, T. S., Álvarez, P. A., León Castro, E., & Uzeta Obregon, C. R. (2024). Analysis of the Countries' business attraction with the ELECTRE-III method. *Computer Science and Information Systems*, *21*(3), 1179-1201. https://doi.org/10.2298/CSIS230223032G
[18] Wu, Q. (2023). Analysis of Advertising Promotion Strategy Based on Improved Collaborative Filtering Algorithm under Digital Media Technology. *International Journal of Computers Communications & Control*, *18*(4), 5392. https://doi.org/10.15837/ijccc.2023.4.5392
[19] Chen, S. C., Lai, M. C., Chu, C. H., Chen, H. M., & Nafei, A. (2024). Enhanced Supplier Evaluation in Digital Transformation: A BWM-Neutrosophic TOPSIS Approach for Decision-Making Under Uncertainty. *Studies in Informatics and Control*, *33*(4), 95-104. https://doi.org/10.24846/v33i4y202409
[20] Wang, S. L. & Zhang, X. (2024). Optimization Strategies and Simulation of Integrated Management in Supply Chains. *International Journal of Simulation Modelling*, *23*(3), 543-554. https://doi.org/10.2507/IJSIMM23-3-CO15
[21] Galluzzo, Y. (2024). A Comprehensive Review of the Data and Knowledge Graphs Approaches in Bioinformatics. *Computer Science and Information Systems*, *21*(3), 1055-1075. https://doi.org/10.2298/CSIS230530027G
[22] Jiang, Z., Chen, Y., Wang, K., Yang, B., & Song, G. (2023). A Graph-Based PPO Approach in Multi-UAV Navigation for Communication Coverage. *International Journal of Computers Communications & Control*, *18*(6), 5505. https://doi.org/10.15837/ijccc.2023.6.5505
[23] Wei, Z. H., Yan, L., & Yan, X. (2024). Optimizing Production with Deep Reinforcement Learning. *International Journal of Simulation Modelling*, *23*(4), 692-703. https://doi.org/10.2507/IJSIMM23-4-CO17
[24] Ribeiro, J., Santos, R., Analide, C., & Silva, F. (2024). Implementing Federated Learning and Explainability Techniques in Regression Models to Increase Transparency and Reliability. *Studies in Informatics and Control*, *33*(4), 15-24. https://doi.org/10.24846/v33i4y202402
[25] Du, H. & Chen, J. (2023). An Improved Ant Colony Algorithm for New Energy Industry Resource Allocation in Cloud Environment. *Tehnički vjesnik-Technical Gazette*, *30*(1), 153-157. https://doi.org/10.17559/TV-20220712164019
[26] Pan, C., Ren, H., Wang, K., Kolb, J. F., Elkashlan, M., Chen, M., & Hanzo, L. (2021). Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions. *IEEE Communications Magazine*, *59*(6), 14-20. https://doi.org/10.1109/MCOM.001.2001076

Contact information:

YiDong ZHU
Dazhou Vocational and Technical College,
Dazhou Sichuan, 635001, China
E-mail: zhuyidong21@126.com