



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

An edge computing system for live-line detection based on multi-sensor data fusion

Yongliang Yao

DOI: [10.1504/IJICT.2026.10076742](https://doi.org/10.1504/IJICT.2026.10076742)

Article History:

Received:	20 November 2025
Last revised:	14 December 2025
Accepted:	18 December 2025
Published online:	17 March 2026

An edge computing system for live-line detection based on multi-sensor data fusion

Yongliang Yao

State Grid Gansu Electric Power Company,
Lanzhou, 730030, China
Email: yaoge1687632@sina.com

Abstract: As smart grid initiatives gain momentum, higher demands have emerged for real-time, precise sensing of power equipment operational status. Traditional live-line detection methods predominantly rely on single sensors, suffering from issues such as low detection efficiency and poor real-time performance. To address these challenges, this paper first designs an edge computing system tailored for live-line data detection in smart grids. Through multi-sensor data fusion, edge nodes can form a collaborative network capable of more rapidly identifying the location of live-line faults. Then, a dual-channel convolutional neural network is designed to accommodate multi-sensor fusion data inputs. A time-frequency attention feature fusion module is incorporated to optimise the feature extraction process. Experimental results demonstrate that the proposed system achieves live detection within 6.3 milliseconds, with a detection accuracy rate of 94.61%. This system enables timely and precise identification of potential equipment faults while maintaining high real-time performance.

Keywords: carbon fibre core conductor; dynamic stress field tracking; multi-head sparse attention mechanism; temporal convolutional network; transformer model.

Reference to this paper should be made as follows: Yao, Y. (2026) 'An edge computing system for live-line detection based on multi-sensor data fusion', *Int. J. Information and Communication Technology*, Vol. 27, No. 22, pp.72–90.

Biographical notes: Yongliang Yao is a Senior Technician at the State Grid Gansu Electric Power Company. He obtained a Bachelor's degree in Electrical Engineering and Automation in Changchun Institute of Technology (June 2010). He published two EI-Index journal papers. His research interests include current transformer equipment, transformers and high voltage tests.

1 Introduction

The electric power sector's rapid growth is propelling not only an ongoing expansion of the grid's scale but also a significant rise in the sheer volume and sophistication of power equipment. Live detection technology, as an important means for power equipment condition monitoring, can obtain real-time operation status information under powered states, detect potential fault risks in a timely manner, effectively avoid power outages caused by equipment failures, reduce economic losses, and improve power supply

reliability (Wan et al., 2018). However, the operating environment of power equipment is complex and variable, and information obtained by a single sensor is often limited, making it difficult to fully and accurately reflect the real operation status of the equipment (Omol et al., 2024). Multisensor data fusion technology can effectively improve the accuracy and reliability of fault detection by comprehensively analysing detection data from multiple different types of sensors and fully utilising complementary information between sensors (Duz et al., 2025). Meanwhile, as the internet of things (IoT) technique rapidly growing, edge computing technology can perform real-time processing and analysis at the source where data is generated, improve system response speed, and meet the strict requirements for real-time performance in smart grid live detection (Chen et al., 2019). Combining multisensor data fusion technology with edge computing to build a live detection edge computing system can not only enable real-time and accurate monitoring of the condition of power equipment but also effectively reduce the computational burden on cloud servers, decrease data transmission costs, and improve the efficiency and reliability of the detection system as a whole (Peng et al., 2019).

In the traditional power industry, live detection mainly involves identifying abnormal conditions in power transmission lines. Since the backend needs to analyse images and visual data from the entire power system, relevant personnel must expend a significant amount of time and effort. Current edge computing combines machine learning with intelligent chips, enabling these devices to have strong computing capabilities while also providing more flexible and efficient decision-making abilities, thus optimising computing nodes (Liang et al., 2024). Huo et al. (2019) collected and pre-processed images using drones. The local server only needs to conduct secondary analysis on the pre-processed images, reducing bandwidth requirements and alleviating communication loads between the frontend and backend. Wu et al. (2023), addressing the issue of high computational demand and insufficient processor power during live detection of transmission line anomalies, mounted the Huawei Atlas edge computing device (Zhao et al., 2024) on drones to achieve high recognition accuracy. Morenas et al. (2023) built edge computing units using the Nvidia TX2 embedded computing device (Dong et al., 2022). The solution proposed by Morenas et al. achieves low-latency, high-reliability, and high-precision processing for transmission line fault detection through edge computing combined with embedded AI hardware, significantly enhancing detection efficiency and system autonomy. This architecture meets the real-time and robustness requirements of smart grids, representing a crucial development direction for fault diagnosis in power systems. Li et al. (2023) established an edge computing platform and designed an edge data processing layer to reduce the time cost during inspections. Ju et al. (2022), presented the design of an edge platform incorporating a specialised data processing layer in their research on mobile edge computing-based image recognition for power equipment inspection. They used image projection to extract inspection image features, significantly reducing energy and time consumption during the inspection process.

Driven by continuous progress in deep learning technology, the field of live detection has undergone profound changes in recent years. These deep learning-based models, leveraging their strong feature extraction capabilities, can automatically extract relevant features from massive raw data and uncover complex nonlinear relationships between inputs and outputs, effectively overcoming the limitations of traditional methods in handling such complex relationships. Pasumponthevar and Jeyaraj (2024) developed an improved faster-convolutional neural network (Faster-CNN) (Liang et al., 2020) model

specifically for detection models in edge computing for power transmission lines. This model uses VGGNet (Guan et al., 2019) as the base network and optimises the detection of power equipment with different aspect ratios and scales by enhancing anchor point generation and non-maximum suppression. Lu et al. (2023) introduced lightweight Ghost convolutions to replace traditional convolutions based on YOLO v5. Furthermore, they used an improved K-means approach to modify the prior bounding boxes, making the optimised model more suitable for deployment on edge computing devices. Gomez et al. (2024) designed an edge computing system for live detection in power systems. By utilising edge computing and multi-sensor perception, they collected power equipment data and designed a residual multi-scale feature fusion CNN model to extract hierarchical features from signals to promote accurate live detection. Wang et al. (2023) collected power equipment fault signals through multi-sensors and converted them into two-dimensional grayscale images. Leveraging CNN's excellent feature extraction capability in image processing and gated recurrent unit (GRU)'s superior feature extraction advantages in handling time series data, they respectively extracted spatial and temporal features from the signals, thereby achieving more precise fault detection. Li et al. (2024) started with the power equipment images themselves perceived through edge computing, and proposed an electrical equipment image recognition method that first performs super-resolution reconstruction and then uses deep learning algorithms for live detection.

Through the research and analysis of live detection for power equipment in existing smart grids, it is evident that the complex fault patterns of power equipment impose clear limitations on traditional single-sensor live detection methods. Moreover, these live detection methods based on signals from a single sensor or features from either the time domain or frequency domain cannot comprehensively reflect the operational status of power equipment, thus failing to meet practical demands. To address the aforementioned challenges, this paper proposes an edge computing system for live detection in terms of multi-sensor data integration. The main work of the system is summarised into the following four aspects.

- 1 An edge computing system for multi-sensor data detection in smart grids was established. This system is architected as a power IoT with a layered stack that includes the application, platform, network, and perception levels. In the edge computing architecture, each sensor node not only collects and performs preliminary processing of local data but also plays a collaborative role during data fusion.
- 2 In the edge computing platform, multi-sensor data fusion enables information sharing and collaborative processing among multiple nodes. In a smart grid, multiple sensors not only collect local data but also exchange information with neighbouring nodes. Through multi-sensor data fusion, edge nodes can form a collaborative network that can more quickly identify the location and type of faults.
- 3 A dual-channel CNN was designed for multi-sensor data feature extraction. By introducing an additional information input channel, a new type of neural network architecture was built. The core idea of this network architecture is to simultaneously process multiple data inputs to achieve feature extraction for complex information. A time-frequency attention feature fusion module and a feature extraction sub-network were designed by combining channel attention, temporal attention, and frequency attention to enhance key fused features.

- 4 After feature fusion of multi-sensor data through time-frequency attention, this paper inputs the fused features into a fully connected layer for live detection of power equipment. The fully linked level applies Softmax regression to gain the classification result for live detection. Experimental outcome indicates that the live detection time of the designed system is reduced by at least 2.2 ms, and the detection accuracy is improved by at least 4.53%. This system can significantly enhance the accuracy of live detection, accurately identify potential equipment faults, and possesses advantages such as high real-time performance, providing strong technical support for the safe and stable operation of power equipment and being of wide application prospects and promotion value.

2 Relevant theory

2.1 Edge computing

Edge computing is an emerging computing paradigm aimed at moving computational and data processing capabilities to the network edge, close to where the data is generated, in order to reduce latency, improve data security and privacy protection, and reduce network bandwidth usage (Alnoman et al., 2019). These distinct advantages position it as a well-suited tool for IoT-based machine signal processing and fault diagnosis. As the IOT develops, 5G communications, and big data technologies, edge computing has garnered growing interest from researchers worldwide. Edge computing offers inherent advantages for power equipment fault detection, including low latency, high bandwidth, and massive connectivity (Kong et al., 2022). Edge computing also plays a key role in machine signal processing, feature extraction, and data fusion, thereby boosting the performance of real-time fault detection (Modupe et al., 2024). Edge nodes possess local computing and decision-making capabilities, enabling them to independently execute critical functions even when disconnected from the cloud. This significantly enhances the overall system's robustness and availability. Furthermore, edge computing does not replace cloud computing; rather, it complements cloud computing through a hybrid architecture of local processing and cloud collaboration. This approach addresses the limitations of cloud computing in real-time performance, bandwidth, privacy, reliability, and scenario adaptability, establishing a division of labour where edge nodes handle real-time tasks while the cloud manages global tasks.

Edge computing integrates core functions of network, computing, storage, and application together to form an open platform (Veeramachaneni, 2025). This platform is located near the data source or the object, which can provide nearby services. By performing computation and generating results near the side where real-time data is generated, the program can respond to the data accordingly, meeting the power industry's requirements for the timeliness, privacy, and security of data. The edge computing solution for IOT devices bridges the information gap between physical equipment and modern industrial management systems. It is worth noting that the cloud computing centre can still store and access historical data generated by the terminal physical equipment.

2.2 Convolutional neural network

CNN uses a local connection method, where each convolution kernel is only connected to a local area of the input data, effectively avoiding the model's overfitting problem. The core architecture of a CNN comprises convolutional level, pooling level, activation functions, and fully linked level (Vonder Haar et al., 2023). The convolutional level scans the image through the convolution kernel to extract local features and achieve effective encoding of spatial information (Zhao et al., 2024b). The pooling level decreases the data dimension through downsampling, decreases the computational load, and simultaneously enhances the model's translation invariance. The fully linked level maps the captured characteristics to the output space for tasks such as classification or regression.

- 1 Acting as the central component of a CNN, the convolutional level is tasked with identifying local features within the input data. It applies convolution operations to the input via a sliding window, extracting features at various positions and producing characteristic maps. The convolution operation is as follows.

$$F[i] = \sum_{k=0}^{K-1} X[i+k] \cdot W[k] \quad (1)$$

where $F[i]$ is the feature value at the i^{th} position, $X[i]$ is the input sequence, $W[k]$ is the weight of the convolution kernel, and K is the size of the convolution kernel.

- 2 The pooling layer serves to downsample feature maps, mitigating computational demands without losing important features. The most common pooling operations are max pooling and average pooling. The output of max pooling is produced by selecting the maximum value in the window, while average pooling derives its output from the mean of all values (Zafar et al., 2022). These operations can retain key features and reduce computational complexity, thereby enhancing the network's robustness to variations in input position.
- 3 As the final stage in a CNN, the fully linked level aggregates features from convolutional and pooling layers for the purpose of final classification or regression (Taye, 2023). The output of the fully linked level is indicated in equation (2).

$$Y = f(W \cdot X + b) \quad (2)$$

where Y is the output vector, W is the weight matrix, X is the flattened feature vector, b is the bias term, and f is the activation function, used for probability normalisation in multi-classification tasks.

3 Edge computing system design for live detection of multi-sensor data in the smart grid

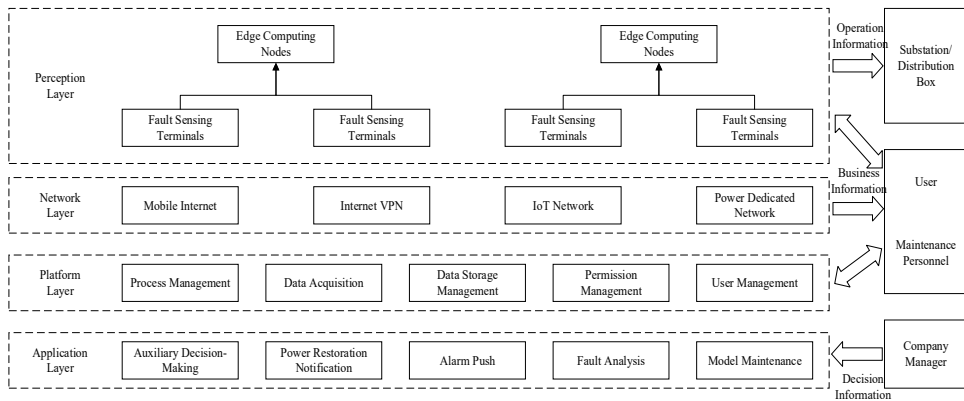
3.1 Overall system architecture

The dominant paradigm in current research is centred on cloud-centric aggregation of data for centralised processing and intelligent applications. However, the rapid increase in terminals poses immense challenges to the cloud computing model, which becomes increasingly inadequate for the massive data transmission, computation, and storage

needs of large-scale intelligent devices. To overcome this limitation, this article introduces an edge computing architecture aimed at multi-sensor data detection within smart grids. The system features a power IoT architecture, consisting of an application level, platform level, network level, and sensing level, as illustrated in Figure 1. The perception layer comprises devices deployed throughout the power distribution system, such as edge computing nodes and multi-sensors. The edge computing nodes and multi-sensors are both embedded hardware devices. The multi-sensors are equipped with adaptive communication interfaces that support various methods to connect to and communicate with the edge computing nodes. Access to the platform layer is gained by the edge computing nodes by means of a network adapter. The primary function of the platform layer is to deploy and operate various service applications, such as those for data storage, log management, permissions, and user management. Oversight and development of the platform layer's application software are handled by the application layer.

Upon collection of live data from smart grid equipment by the multi-sensors, the edge computing model is employed to transmit this data to the cloud server. Upon dividing the collected data into metadata and residual data, the edge server proceeds to encrypt both using the advanced encryption standard (AES). Following encryption, the data is transmitted to the residual data block within the cloud server. This process employs the binary Reed-Solomon coding principle (Retter, 2002) to handle the metadata gathered from the sensors. The cloud server stores the redundant data blocks, along with any majority data blocks that have a count less than k . The edge computing model sends the synchronised test results to the user upon completion of its calculations, enabling the download of any required data. The cloud forwards the majority data blocks and residual data blocks to the edge layer in the edge computing model. Simultaneously, the server transmits the critical metadata blocks to the edge layer, which, upon receipt, decrypts them along with the residual blocks. Following decryption, the data undergoes decoding, after which the processed test data is returned to the local user.

Figure 1 Edge computing system architecture for live-line inspection in the smart grid



3.2 Multi-sensor data fusion based on the edge computing platform

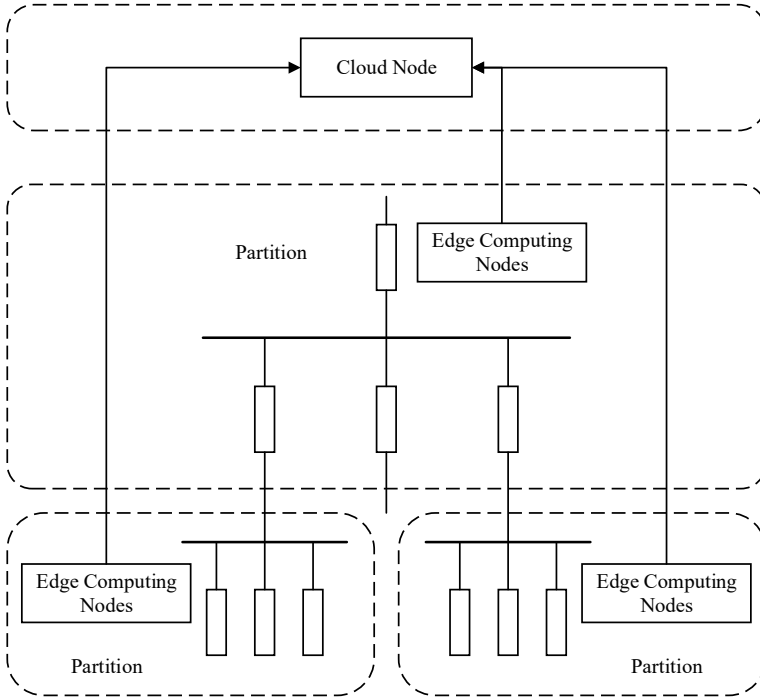
The voltage, current, power, and other data collected by each sensor node in the smart grid contain noise, redundancy, or partial loss. In the edge computing architecture, each sensing node is not only responsible for the collection and preliminary processing of local data but also plays a collaborative role in the data fusion process, as indicated in Figure 2. After the edge computing platform receives data from different nodes, it performs local computation to clean, correct, and integrate the data, reducing redundancy and errors, and improving the accuracy of live-line detection.

Multiple nodes in the smart grid simultaneously collect similar current and voltage data, which may deviate due to sensor accuracy, noise during transmission, and other factors. Thus, this paper adopts the weighted average method for data fusion. Suppose current data $I_i(t)$ is collected at node i in the smart grid, where t is time and the weight w_i of the node can be set based on the sensor's reliability or the criticality of its location, then the weighted average current value $I_{avg}(t)$ is as follows.

$$I_{avg}(t) = \frac{\sum_{i=1}^n \omega_i I_i(t)}{\sum_{i=1}^n \omega_i} \quad (3)$$

where n is the number of nodes involved in fusion, and w_i is the weight of node i . In this way, the edge computing node can eliminate the influence of unreliable data and obtain a more accurate current measurement.

Figure 2 Collaborative processing in edge computing architectures



Meanwhile, this paper utilises the Kalman filter (Zhao and Hu, 2023) for real-time estimation of dynamic systems. The Kalman filter recursively fuses observation data from multiple sensors to estimate system states. Let the true current value at a node in the smart grid be $I_{true}(t)$, and the sensor measurement value be $I_{measured}(t)$. Then the system state is estimated by the Kalman filter, with the result as shown in equation (4).

$$\hat{I}_{true}(t) = \hat{I}_{true}(t-1) + K(t)[I_{measured}(t) - \hat{I}_{true}(t-1)] \quad (4)$$

where $\hat{I}_{true}(t)$ is the estimated current value, and $K(t)$ is the Kalman gain, reflecting the size and weight of measurement errors. At each moment, the Kalman filter updates based on the current measurement value and the previous estimate, thereby providing a more accurate current estimate. Through the Kalman filter, measurement noise can be effectively eliminated, and precise current and voltage values can be provided for subsequent fault detection.

On the edge computing platform, information sharing and collaborative processing among multiple nodes are achieved through multi-sensor data fusion. In the distribution network, multiple sensors not only collect local data but can also exchange information with neighbouring nodes. Through multi-sensor data fusion, edge nodes can form a collaborative network that can more quickly identify the location and type of a fault. When the current value at a certain node exceeds the set threshold, the edge node sends an alarm message to the neighbouring nodes. After obtaining voltage, current, and other data from surrounding nodes, comprehensive analysis is performed. If multiple nodes show consistent anomalies, the fault location can be further confirmed.

4 Inference of tourist behaviour and consumption intentions based on the CNViT algorithm

4.1 Structure of the model for inferring tourist behaviour and consumption intentions

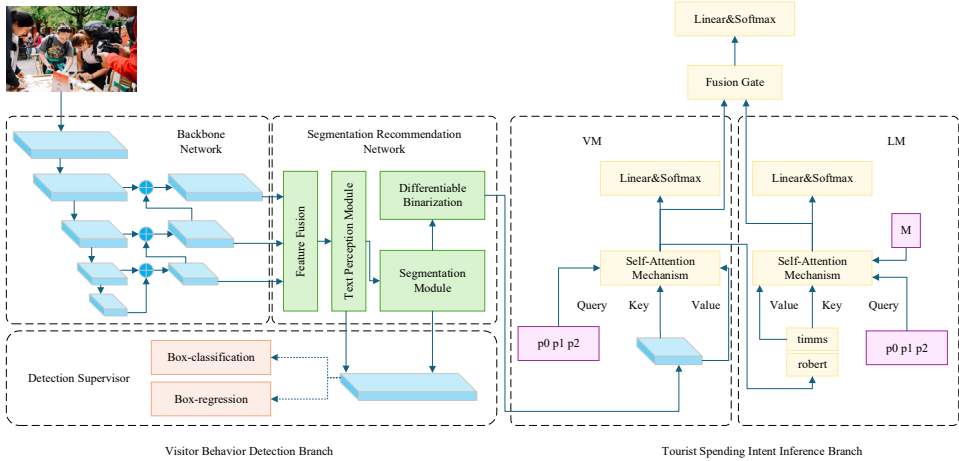
Data for inferring tourist behaviour and purchase intent is primarily found on social media platforms, online travel agencies, and travel forums. This paper designs a tourist behaviour and purchase intent inference model based on the CNViT algorithm, as shown in Figure 3. The model comprises three key components that allow for the accurate inference of both tourist behaviour and purchase intent.

- 1 **Backbone network.** It uses CNN as its backbone to extract rich multi-level visual features from tourist behaviour image. Not only these basic visual features include colour, texture, shape, but also these visual features contain rich semantic information. They provide very important clues for tourist purchase intent inference. Main network will extract features step by step through multiple layers convolution, deeply mining the image to get the deep features. They are fine-grained features that can deeply distinguish text from background and text from non-text objects. It makes the whole subsequent recognition possible.
- 2 **Tourist behaviour detection branch.** It is used to accurately locate the text region in the image of tourist behaviour. It contains two main sub-modules: segmentation network and detection supervisor. The segmentation network uses advanced deep

learning segmentation method to divide the image finely, and accurately find the candidate area that may contain the text. The detection supervisor will play a very important role in the training process to provide supervisory signal for the detection.

- 3 Purchase intent inference branch. It uses the visual module (VM) and language module (LM) of CNViT algorithm and add a fusion gate to deeply understand and recognise the text content. VM module refines and improves the initial text area output from detection branch. LM module uses self-attention method to extract contextual dependencies of text and accurately segment the semantic information of text. Finally, the fusion gate deeply combines the feature from VM and LM module and extract the most valuable feature representation. It provides the final tourist purchase intent inference with comprehensive and complementary feature information.

Figure 3 The suggested tourist behaviour and purchase intent inference model (see online version for colours)



4.2 Backbone network

As shown in Figure 3, the main network is located in the centre of our model. Its task is to extract rich bottom-up multi-level visual features from input tourist images. These visual features can represent fine-grained texture, shape and semantic information of input images, and provide basis for next purchase intent inference. In this paper, we select feature pyramid network net (FPNNet) (Liu et al., 2023) as main network to extract key visual features from input tourist images. In the process of down-sampling, we can get spatial attention module by utilising spatial relationship of intermediate text feature information. In each convolution, a feature information matrix I is generated, and the sigmoid function is applied to it for activation, as shown in equation (3).

$$W_s(I) = \sigma^{3 \times 3} Pool(I) \quad (5)$$

$$W_s(I) = \sigma^{3 \times 3} Pool(I) \quad (6)$$

where $f^{3 \times 3}$ is the convolution operation, with a 3×3 convolutional kernel. During the up-sampling process, we use the uppool pooling method to extract features and generate a channel attention module to approximate the location features. Then, the features are adjusted through a shared MLP, as shown in equation (4).

$$W_c(I') = \sigma \text{MLP}(\text{unpool}(I)) = \sigma W_1 W_0 I' \quad (7)$$

where σ is the sigmoid activation function, and W_0 and W_1 are the weights of the MLP. Finally, during the feature fusion stage, the weights of the channel attention and spatial attention are combined into an attention model to enhance the feature representation capability. This process is represented as follows.

$$I = (W_s(I) + 1) \odot I \quad (8)$$

$$I'' = (W_c(I') + 1) \odot I' \quad (9)$$

where $\odot$ is the element-wise multiplication of the corresponding matrix elements. Since the sigmoid function is used to activate the elements of the attention channel, the attention module can better enhance useful image information and suppress useless information by limiting the values of the attention channel elements. Specifically, by restricting each element's value to the range $[0, 1]$, this effect can be achieved.

4.3 Visitor behaviour detection branch

Detection branch applies convolutional module of UNet (Trebing et al., 2021) to integrate and process multi-scale visual features extracted from backbone network. The purpose of doing this is to enhance the ability of model to detect text of different sizes and proportions, and then improve its scale robustness. Specifically, for the three kinds of feature maps that are extracted from FPNNet backbone network and have different level and receptive field, we apply 3×3 convolutional layer on each of them. Then, through upsampling operations, these features are restored to the same size, achieving scale normalisation. After the above processing, the features at each scale are concatenated together to form a multi-scale fusion feature. The fused feature can be represented as follows.

$$F = \text{UNet}(R_v) \quad (10)$$

To further improve the model's accuracy in detecting tourist behaviour, a text-aware module is designed and introduced after the visual feature fusion module of the backbone network. This module is specifically optimised for text information. The fused feature, after passing through the text-aware module, becomes the text-aware feature, denoted as $F_{tan} = T(F)$.

After applying the text-aware module to extract F_{tan} , this paper further refines the text feature representation. Based on this, we can define the feature map after image segmentation processing as $S = S(F_{tan})$, where the last layer of the segmentation module is a sigmoid layer, meaning the range of S is $[0, 1]$.

Considering that text in tourist behaviour images is often blurry, making it difficult to distinguish between text and non-text regions, this paper proposes a separable binarisation formula named attention-DBNet, which incorporates threshold selection

during the binarisation process into the model training, allowing the model to adaptively select the threshold. Binarising the segmentation feature map S , the binarised segmentation result is as follows, where $B_{i,j}$ is the binarised result at position (i, j) , x is the value at position (i, j) of the segmentation map, and k is the magnification factor.

$$B_{i,j} = \frac{1}{1 + e^{\frac{e^{-kx} - e^{kx}}{2}}} \quad (11)$$

This paper uses the cross-entropy loss function to explain why the proposed attention-DBNet binarisation formula performs better on target images. Using equation (11) as the original function, the positive label sample loss function I_+ is shown in equation (12), and the negative label sample loss function I_- is shown in equation (13).

$$I_+ = -\ln \frac{1}{1 + e^{\frac{e^{-kx} - e^{kx}}{2}}} \quad (12)$$

$$I_- = -\ln \left(1 - \frac{1}{1 + e^{\frac{e^{-kx} - e^{kx}}{2}}} \right) \quad (13)$$

Inspired by Fang et al. (2019), in the model training phase of this study, the fast R-CNN architecture is used as a detection supervisor for the tourist behaviour detection module in the training process. In this way, fast R-CNN takes on the role of accurately detecting performance supervision for the segmentation recommendation network, thereby enhancing the overall system's detection capability.

4.4 Branch for inferring tourist consumption intentions

In our proposed tourist behaviour and consumption intention inference model, the whole process is modelled as directly continuous from image visual features to consumption intention text semantics. While dealing with task of inferring and recognising tourists' consumption intentions, the position information of character sequence is of great importance. On one hand, it represents the order of appearance of characters; on the other hand, it also contains a lot of semantic clues. In order to make model fully use and capture the influence of position information, integrate-dimensional positional encoding into sequence decoding process based on CNViT is an effective way to improve accuracy of recognition. This paper uses the sine and cosine functions to encode positional information, which can effectively encode the position of characters in the sequence in a continuous and periodic manner, thereby enabling the model to distinguish characters at different positions, even if they have the same character content, different representations can be assigned due to positional differences. Positional encoding can be represented as follows, where pos represents the position index of a character in the text sequence, with its odd-dimensional index in the position vector being $2 + 1$ and its even-dimensional index being 2, dim represents the dimension of the position vector.

$$PE_{(pos, 2i)} = \sin \left(\frac{pos}{10000^{\frac{2i}{dim}}} \right) \quad (14)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{\frac{2i}{10000^{dim}}}\right) \quad (15)$$

In the design of the VM module in the tourist consumption intention inference branch, the positional encoding results obtained from equations (11) and (12) are first used as the query input for CNViT. These positional encodings reflect the position information of characters in the text sequence. Meanwhile, the text visual features F_{tan} extracted by the text perception module are used as the key and value for CNViT. Therefore, CNViT can perform decoding operations based on the correlation between positional query information and text visual features, thus extracting high-level semantic features corresponding to each character. Finally, the text character features decoded by the VM module can be represented as follows.

$$g_v = \text{softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \quad (16)$$

where Q , K , and V are the query, key, and value, respectively, and K^T is the transpose of K . To fully comprehend and process the complex interaction information between text visual features and characters, we apply Transformer module with 8 parallel attention heads. Each attention head independently capture and learn the correlation in input sequence from its own view, so as to achieve multi-dimensional understanding and parsing of characters' multi-faceted dependencies. Then, the decoded text character features g_v are processed through a linear transformation and SoftMax activation function to generate the final recognition probability distribution, where P_v is the recognition result of the VM, and $T_v = W_v(g_v)$, W_v are trainable hyperparameters.

$$P_v = \text{softmax}(T_v) \quad (17)$$

The LM module is very important in tourist consumption intention inference task. Its basic role is to further deeply mine and utilise the initial results obtained from visual recognition process. Inspired by the view in Koo et al. (2016), we regard tourist consumption intention inference process as a complex bidirectional understanding and fill-in-blank process, just like the bidirectional masked language model task of predicting obscured words. Through this kind of structural design, we can make use of advantages of positional encoding to extract and integrate semantic content that is most related to context within the already extracted and encoded visual features that have been extracted and encoded, and then achieve accurate inference of tourist consumption intentions.

The self-attention module in LM module applies a specially customised mask matrix M . The design of masking strategy is to simulate the operational situation of bidirectional fill-in-the-blank. In the process of encoding, make sure that the representation of current character is not depended on direct information of it, but constructed entirely by clues from context information of surrounding characters. The encoding process is represented by equation (18), where Q is the positional encoding of the character, and K and V are the key and value, respectively. The text recognition result obtained after processing by the LM module can be expressed by equation (19), where P_L is the recognition output of the LM, and $F_L = W_l(g_L)$, W_l are trainable hyperparameters.

$$g_L = \text{softmax}\left(\frac{QK^T}{\sqrt{C}} + M\right)V \quad (18)$$

$$P_L = \text{softmax}(F_L) \quad (19)$$

To effectively integrate visual features with semantic features to enhance the performance of tourist consumption intention inference, a fusion gate mechanism is set up in the model. First, the g_v encoded by the VM module and the semantic features g_L extracted by the LM module are input into the fusion gate for integration, as shown in equation (20).

$$\begin{cases} W_f = \sigma([g_v, g_L]W_g) \\ F_g = W_f g_v + (1 - W_f) g_L \end{cases} \quad (20)$$

where σ is the sigmoid activation function, used to generate a weight value between 0 and 1, which controls the relative contributions of g_v and g_L in the fusion process. $[g_v, g_L]$ denotes the concatenation of the two along a dimension to form a new feature vector, and W_g and W_f represent trainable hyperparameters and fusion weights, respectively. The prediction result of tourist consumption intention inference can be expressed as equation (21).

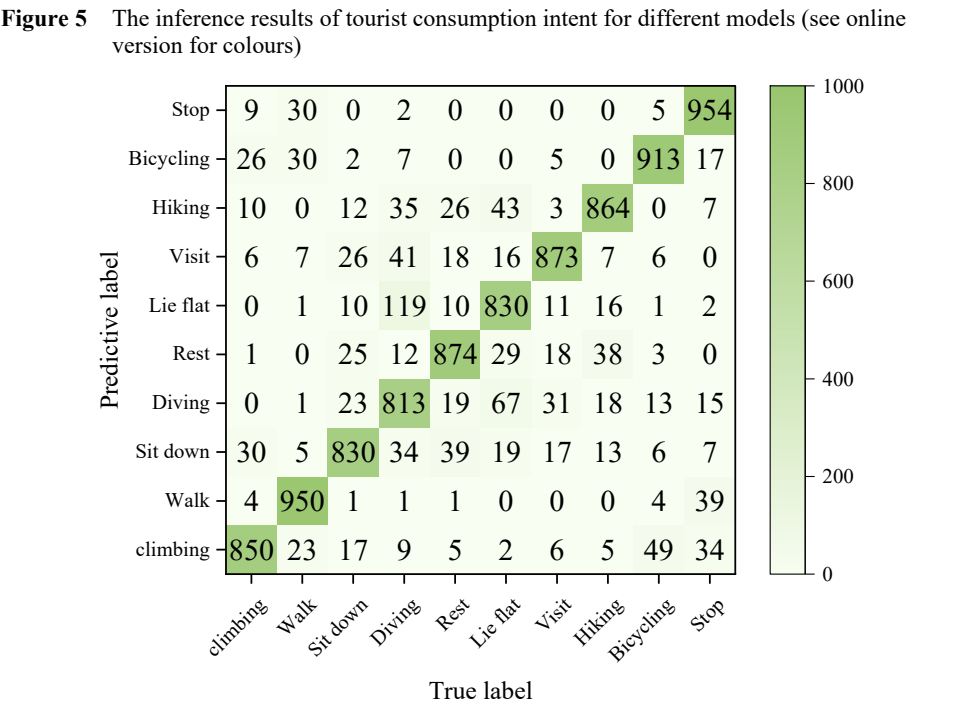
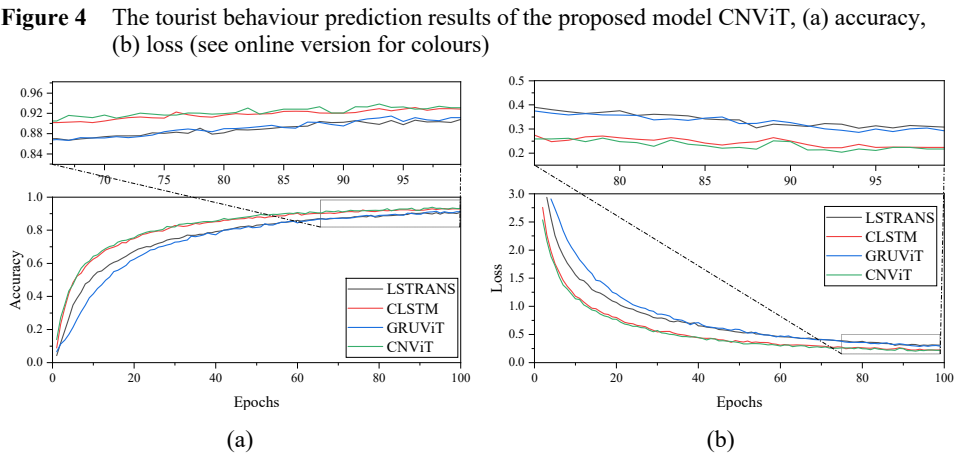
$$P_F = \text{softmax}(F_f) \quad (21)$$

where P_F is the predicted probability distribution output by the tourist consumption intention inference branch, and F_f is the fused feature after linear transformation, specifically $F_f = W_r(F_g)$, and W_r are trainable hyperparameters. Through this series of operations, the model can flexibly adjust the weights of visual and semantic information based on actual conditions, thus achieving more accurate and comprehensive tourist consumption intention inference.

5 Experimental results and performance analysis

5.1 Analysis of tourist behaviour and inference results on their consumption intentions

This paper uses the tourist behaviour and consumption data collected in reference (Li and Cao, 2022) as the experimental dataset. This dataset includes various information on tourist behaviour from 20,563 records on a travel website, such as UserID, the average number of times users browse travel-related pages annually, user reviews of travel destinations, and hotel consumption data, etc. The dataset is divided into training, testing, and validation sets at a ratio of 6:2:2. The AdamW algorithm with an initial learning rate of 0.0001 is chosen for gradient training, using a learning rate decay strategy where the learning rate decays to half of its original value every 50 epochs, with a maximum of 200 training epochs. The experiment uses an NVIDIA GeForce RTX 4060 Laptop GPU processor, and the program runs in the PyCharm software under the PyTorch 2.0.1 framework, with Python3 as the programming language.⁹ is the programming language.



The tourist behaviour prediction results of the proposed model CNViT are shown in Figure 4, where the x-axis represents the true labels of the model, the y-axis represents the predicted labels, and the value in each cell represents the number of samples of a certain category that the model predicts into another category. From Figure 4, it can be seen that the model performs relatively accurately in most categories, especially in the ‘walk’ and ‘stop’ categories, where 950 and 954 samples, respectively, were correctly classified, indicating very high classification accuracy in these two categories. However, the model has some confusion in certain categories, for example, 119 samples between

the ‘sit down’ and ‘rest’ categories were misclassified, indicating that the model has some difficulty in distinguishing between ‘sit down’ and ‘rest’.

For ease of analysis, LSTRANS (Liu and Hu, 2025), CLSTM (Shao and Kim, 2020), and GRUViT (Si, 2025) were selected as benchmark models. The inference results of tourist consumption intent for different models are shown in Figure 5. The CNViT model demonstrated the fastest convergence speed during training, achieving the highest accuracy in inferring tourist consumption intent and the lowest loss value. The accuracy of GRUViT was slightly lower than CNViT, while the loss value was slightly higher. In comparison, the LSTRANS model had the slowest convergence rate, followed by the CLSTM model.

5.2 Performance analysis of different methods

The prediction accuracy of CNViT and GRUViT for various tourist behaviours is shown in Figure 6. The average prediction accuracy of CNViT for various tourist behaviours reached 96.17%, while the average prediction accuracy of GRUViT for various tourist behaviours was 95.03%. The average prediction accuracy of CNViT was 1.14% higher than that of GRUViT. Although GRUViT uses a hybrid model of GRU and ViT for predicting tourist behaviour, the GRU mitigates gradient vanishing in long sequences through gating mechanisms, but still struggles to capture ultra-long dependencies in tourists’ continuous behaviours spanning several hours (such as travel routes and consumption decisions). Additionally, when GRUViT performs global self-attention calculations on tourist behaviour images, it may focus on irrelevant regions, increasing computational costs and reducing feature effectiveness. CNViT uses Unet to achieve the detection of tourist behaviour and designs a ViT segmentation network to finely divide images, accurately detecting tourist behaviour and improving the prediction accuracy of tourist behaviour.

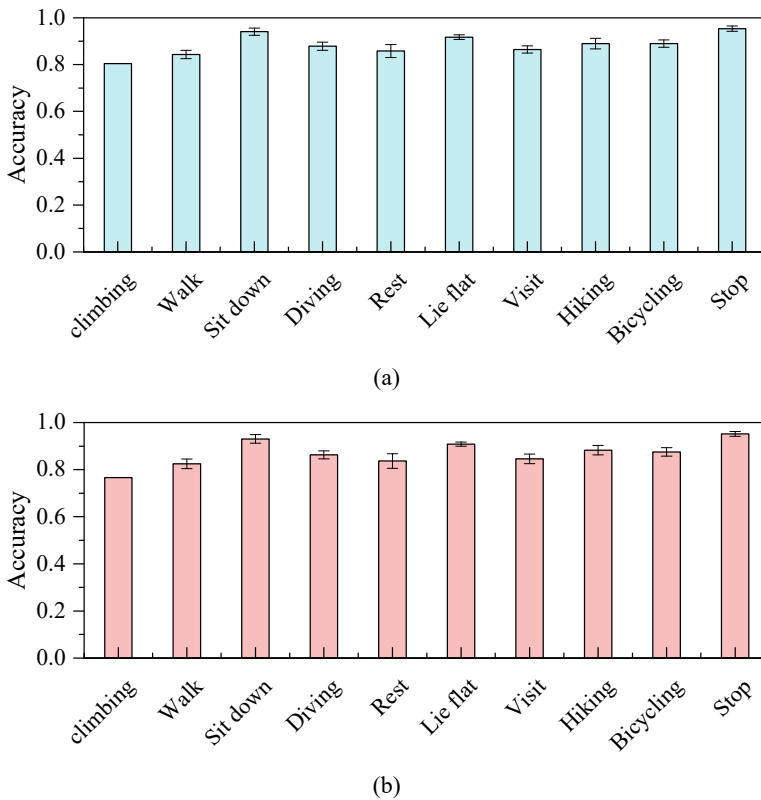
Table 1 Comparison of the inference performance for various methods

Method	Tourist behaviour prediction				Consumption intention inference			
	A	R	F1	AUC	A	R	F1	AUC
LSTRANS	0.875	0.842	0.857	0.902	0.839	0.846	0.839	0.898
CLSTM	0.906	0.886	0.884	0.948	0.887	0.871	0.891	0.947
GRUViT	0.941	0.905	0.931	0.975	0.904	0.918	0.928	0.961
CNViT	0.968	0.937	0.959	0.983	0.942	0.935	0.963	0.979

This paper further analyses the performance of LSTRANS, CLSTM, GRUViT, and CNViT in inferring tourist behaviour and consumption intentions using evaluation metrics such as accuracy (A), recall (R), F1, and AUC, as shown in Table 1. The accuracy of CNViT in inferring tourist behaviour and consumption intentions is 96.8% and 94.2%, respectively, which is an improvement of 9.3% and 10.3% over LSTRANS, 6.2% and 5.5% over CLSTM, and 2.7% and 3.8% over GRUViT. Comparing the harmonic mean of precision and recall, F1, CNViT’s F1 scores for inferring tourist behaviour and consumption intentions are 95.9% and 96.3%, respectively, which is an improvement of 10.2% and 12.4% over LSTRANS, 7.5% and 7.2% over CLSTM, and 2.8% and 3.5% over GRUViT. Comparing the area under the ROC curve (AUC), CNViT’s AUC scores for inferring tourist behaviour and consumption intentions are

0.983 and 0.979, respectively, which is at least an improvement of 0.82% and 1.87% over the other three models. LSTRANS mitigates the vanishing gradient problem of traditional RNNs through gating mechanisms, but may still fail to capture long-term dependencies due to gradient decay when processing long-term tourist behaviour trajectories. Additionally, ViT captures global information through the self-attention mechanism but may lack sufficient extraction of local details in tourist consumption scenarios, potentially affecting the judgment of consumption intentions. CLSTM typically simple concatenates the spatial features extracted by CNN and the temporal features captured by LSTM through a fully connected layer, lacking a deep feature interaction mechanism. Additionally, it may ignore the correlation between spatial and temporal features, leading to inference errors. GRUViT, which infers tourist behaviour and consumption intentions based on GRU and ViT, but this method ignores the visual features of tourist text reviews, resulting in low inference accuracy. CNViT combines the local features extracted by CNN with the global features captured by ViT, allowing for more accurate inference of consumption intentions.

Figure 6 The prediction accuracy of, (a) CNViT and (b) GRUViT for various tourist behaviours (see online version for colours)



6 Conclusions

With the rapid development of tourism industry, large scale of tourist images and behaviour data open up new opportunities to accurately understand tourist's real preferences and consumption intentions. However, traditional methods ignore the spatiotemporal complexity of tourist behaviour and implicit semantic associations of tourist's consumption intentions, which cause the low inference accuracy. In this paper, we first optimise the ViT algorithm by integrating with hierarchical structure of CNNs. Then, we design tourist behaviour and consumption intention inference model based on optimised ViT algorithm. The two branches of tourist behaviour detection and consumption intention inference make the whole model accurately extract deep semantic information from tourist's images and efficiently decode and understand tourist's review text, which make our model get excellent performance in tourist consumption intention inference. Specifically, the consumption intention inference branch adopts the VM and LM modules as well as fusion gate component. The whole model fully extract visual features and further explore semantic information of text. Through positional encoding, VM and LM can recognition operation together, which make the model get the advantage of rapid decoding, thus achieve accurate consumption intention inference. The experimental results on real datasets show that compared with the baseline models, the AUC of tourist behaviour and consumption intention inference achieve at least 2.8% and 3.5% improvement respectively, which demonstrate the excellent inference performance.

Although the model can achieve the accurate inference, most of current research still depends on textual and visual data. Tourist behaviour is influenced by multi-modal information such as voice interaction and environmental perception. In the future, a multi-modal pre-training framework will be constructed to uniformly model text features, visual features, voice features, and environmental sensor data through a cross-modal Transformer, enhancing the robustness of tourist behaviour and consumption intention inference.

Declarations

All authors declare that they have no conflicts of interest.

References

- Alnoman, A., Sharma, S.K., Ejaz, W. and Anpalagan, A. (2019) 'Emerging edge computing technologies for distributed IoT systems', *IEEE Network*, Vol. 33, No. 6, pp.140–147.
- Chen, S., Wen, H., Wu, J., Lei, W., Hou, W., Liu, W., Xu, A. and Jiang, Y. (2019) 'Internet of things based smart grids supported by intelligent edge computing', *IEEE Access*, Vol. 7, No. 5, pp.74089–74102.
- Dong, J., Zheng, F., Lin, J., Liu, Z., Xiao, F. and Fan, G. (2022) 'EC-ECC: accelerating elliptic curve cryptography for edge computing on embedded GPU TX2', *ACM Transactions on Embedded Computing Systems (TECS)*, Vol. 21, No. 2, pp.1–25.
- Duz, F.H.d.S., Zacarias, T.G., Ribeiro Junior, R.F., Steiner, F.M., Assuncao, F.d.O., Bonaldi, E.L. and Borges-da-Silva, L.E. (2025) 'Smart monitoring of power transformers in Substation 4.0: multi-sensor integration and machine learning approach', *Sensors*, Vol. 25, No. 17, pp.54–69.

- Fang, F., Li, L., Zhu, H. and Lim, J-H. (2019) 'Combining faster R-CNN and model-driven clustering for elongated object detection', *IEEE Transactions on Image Processing*, Vol. 29, No. 2, pp.2052–2065.
- Gomez, P.I., Gajardo, M.E.L., Mijatovic, N. and Dragičević, T. (2024) 'A self-commissioning edge computing method for data-driven anomaly detection in power electronic systems', *IEEE Transactions on Industrial Electronics*, Vol. 71, No. 10, pp.13319–13330.
- Guan, Q., Wang, Y., Ping, B., Li, D., Du, J., Qin, Y., Lu, H., Wan, X. and Xiang, J. (2019) 'Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study', *Journal of Cancer*, Vol. 10, No. 20, pp.48–56.
- Huo, W., Liu, F., Wang, L., Jin, Y. and Wang, L. (2019) 'Research on distributed power distribution fault detection based on edge computing', *IEEE Access*, Vol. 8, No. 2, pp.24643–24652.
- Ju, X., Gou, R., Xiao, Y., Wang, Z. and Liu, S. (2022) 'The use of edge computing-based internet of things big data in the design of power intelligent management and control platform', *International Journal of Grid and Utility Computing*, Vol. 13, No. 1, pp.76–86.
- Kong, X., Wu, Y., Wang, H. and Xia, F. (2022) 'Edge computing for internet of everything: a survey', *IEEE Internet of Things Journal*, Vol. 9, No. 23, pp.23472–23485.
- Koo, C., Joun, Y., Han, H. and Chung, N. (2016) 'A structural model for destination travel intention as a media exposure: belief-desire-intention model perspective', *International Journal of Contemporary Hospitality Management*, Vol. 28, No. 7, pp.1338–1360.
- Li, D., Lu, Y., Gao, Q., Li, X., Yu, X. and Song, Y. (2024) 'LiteYOLO-ID: a lightweight object detection network for insulator defect detection', *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, No. 4, pp.1–12.
- Li, J. and Cao, B. (2022) 'Study on tourism consumer behavior and countermeasures based on big data', *Computational Intelligence and Neuroscience*, Vol. 4, No. 1, pp.61–78.
- Li, Y., Liu, M., Li, Z. and Jiang, X. (2023) 'CSSAdet: real-time end-to-end small object detection for power transmission line inspection', *IEEE Transactions on Power Delivery*, Vol. 38, No. 6, pp.4432–4442.
- Liang, H., Zuo, C. and Wei, W. (2020) 'Detection and evaluation method of transmission line defects based on deep learning', *IEEE Access*, Vol. 8, pp.38448–38458.
- Liang, S., Jin, S. and Chen, Y. (2024) 'A review of edge computing technology and its applications in power systems', *Energies*, Vol. 17, No. 13, pp.32–45.
- Liu, N. and Hu, D. (2025) 'The design of consumer behavior prediction and optimization model by integrating DQN and LSTM', *Plos One*, Vol. 20, No. 7, pp.32–48.
- Liu, W., Zhou, B., Wang, Z., Yu, G. and Yang, S. (2023) 'FPPNet: a fixed-perspective-perception module for small object detection based on background difference', *IEEE Sensors Journal*, Vol. 23, No. 10, pp.11057–11069.
- Lu, L., Chen, Z., Wang, R., Liu, L. and Chi, H. (2023) 'Yolo-inspection: defect detection method for power transmission lines based on enhanced YOLOv5s', *Journal of Real-Time Image Processing*, Vol. 20, No. 5, pp.104–112.
- Modupe, O.T., Otitoola, A.A., Oladapo, O.J., Abiona, O.O., Oyeniran, O.C., Adewusi, A.O., Komolafe, A.M. and Obijuru, A. (2024) 'Reviewing the transformational impact of edge computing on real-time data processing and analytics', *Computer Science & IT Research Journal*, Vol. 5, No. 3, pp.693–702.
- Morenas, J., Moya-Fernández, F. and López-Gómez, J.A. (2023) 'The edge application of machine learning techniques for fault diagnosis in electrical machines', *Sensors*, Vol. 23, No. 5, pp.26–49.
- Omol, E., Mburu, L. and Onyango, D. (2024) 'Anomaly detection in IoT sensor data using machine learning techniques for predictive maintenance in smart grids', *International Journal of Science, Technology & Management*, Vol. 5, No. 1, pp.201–210.

- Pasumponthevar, M.K. and Jeyaraj, P.R. (2024) 'Reliability analysis and Terahertz characterization of circuit board defect employing branchy neural network in edge-cloud fusion architecture', *Engineering Applications of Artificial Intelligence*, Vol. 138, No. 6, pp.109–118.
- Peng, Y., Tan, A., Wu, J. and Bi, Y. (2019) 'Hierarchical edge computing: a novel multi-source multi-dimensional data anomaly detection scheme for industrial Internet of Things', *IEEE Access*, Vol. 7, No. 2, pp.111257–111270.
- Retter, C.T. (2002) 'Gaps in the binary weight distributions of Reed-Solomon codes', *IEEE Transactions on Information Theory*, Vol. 38, No. 6, pp.1688–1697.
- Shao, X. and Kim, C.S. (2020) 'Multi-step short-term power consumption forecasting using multi-channel LSTM with time location considering customer behavior', *IEEE Access*, Vol. 8, No. 1, pp.125263–125273.
- Si, W. (2025) 'Application of deep learning in consumer purchase intention prediction', *Service Oriented Computing and Applications*, Vol. 7, No. 5, pp.1–15.
- Taye, M.M. (2023) 'Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions', *Computation*, Vol. 11, No. 3, pp.52–67.
- Trebing, K., Stańczyk, T. and Mehrkanoon, S. (2021) 'SmaAt-UNet: precipitation nowcasting using a small attention-UNet architecture', *Pattern Recognition Letters*, Vol. 145, No. 4, pp.178–186.
- Veeramachaneni, V. (2025) 'Edge computing: Architecture, applications, and future challenges in a decentralized era', *Recent Trends in Computer Graphics and Multimedia Technology*, Vol. 7, No. 1, pp.8–23.
- Vonder Haar, L., Elvira, T. and Ochoa, O. (2023) 'An analysis of explainability methods for convolutional neural networks', *Engineering Applications of Artificial Intelligence*, Vol. 117, No. 10, pp.16–28.
- Wan, Z., Li, J. and Gao, Y. (2018) 'Monitoring and diagnosis process of abnormal consumption on smart power grid', *Neural Computing and Applications*, Vol. 30, No. 1, pp.21–28.
- Wang, Y., Liao, Y., Yang, J., Wang, H., Zhao, Y., Zhang, C., Xiao, B., Xu, F., Gao, Y. and Xu, M. (2023) 'An FPGA-based online reconfigurable CNN edge computing device for object detection', *Microelectronics Journal*, Vol. 137, No. 10, pp.105–118.
- Wu, S., Wang, L., Song, B. and Gao, J. (2023) 'Live-line detection of deteriorated insulators on overhead transmission lines based on electric field distribution', *IEEE Sensors Journal*, Vol. 24, No. 2, pp.1687–1695.
- Zafar, A., Aamir, M., Mohd Nawi, N., Arshad, A., Riaz, S., Alruban, A., Dutta, A.K. and Almotairi, S. (2022) 'A comparison of pooling methods for convolutional neural networks', *Applied Sciences*, Vol. 12, No. 17, pp.86–93.
- Zhao, B., Qin, Z., Wu, Y., Song, Y., Yu, H. and Gao, L. (2024a) 'A fast target detection model for remote sensing images leveraging roofline analysis on edge computing devices', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 6, No. 4, pp.72–80.
- Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M. and Parmar, M. (2024b) 'A review of convolutional neural networks in computer vision', *Artificial Intelligence Review*, Vol. 57, No. 4, pp.99–116.
- Zhao, H. and Hu, J. (2023) 'Iterative unscented Kalman filter with general robust loss function for power system forecasting-aided state estimation', *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, No. 6, pp.1–9.