https://doi.org/10.62836/jcmea.v6i1.0001

Global Science Publishing

JCMEA

Journal of Computational Methods in Engineering Applications

https://ojs.sgsci.org/journals/jcmea

Article

Medical Biopharmaceutical Image Anomaly Detection under Retinex State Space Duality and Frequency Consensus-Driven Transformer

Yuhan Dai

Biostats & Data Management, Arrowhead Pharmaceuticals, Pasadena, CA 91105, USA

Abstract: With the rapid development of medical imaging technologies and biopharmaceutical research, a large amount of high-dimensional and complex medical and biological image data is continuously being generated. Achieving high-precision anomaly detection under conditions of complex backgrounds and low contrast has become an important research problem in the fields of intelligent healthcare and drug development. Traditional anomaly detection methods often struggle to achieve stable and robust detection performance when dealing with issues commonly present in medical images, such as uneven illumination, complex structures, and insufficient utilization of frequency information. To address these challenges, this paper proposes a Transformer-based anomaly detection method for medical and biopharmaceutical images driven by Retinex state-space duality and frequency consensus. By integrating spatial-domain and frequency-domain features, the proposed method enhances the model's ability to perceive complex abnormal structures. Specifically, the method first employs Retinex state-space duality to decompose medical images into structural and illumination components, thereby strengthening the structural representation of anomalous regions while reducing interference caused by illumination variations. Subsequently, a frequency consensus-driven mechanism is introduced to model feature consistency across different scales in the frequency domain, enabling adaptive enhancement of key anomalous frequency features. On this basis, a global context modeling framework is constructed by incorporating a Vision Transformer, which captures long-range dependencies and potential anomalous patterns in medical images, further improving detection accuracy and feature representation capability. To verify the effectiveness of the proposed method, extensive experiments are conducted on multiple medical and biopharmaceutical image datasets, and comparisons are made with several mainstream anomaly detection models. Experimental results demonstrate that the proposed method outperforms the comparison methods across evaluation metrics, achieving more stable and accurate anomaly detection in complex medical imaging environments. This approach provides a technically promising solution with potential application value for automated image analysis in intelligent pharmaceutical inspection and drug development.

Keywords: Vision Transformer; Retinex state-space duality; frequency consensus-driven; anomaly detection; data analysis

1. Introduction

With the continuous development of medical imaging technology and biopharmaceutical research, a large number of pharmaceutical images and biological microscopic images are playing an increasingly important role in disease

Received: 15 March 2026; Accepted: 30 March 2026.

Corresponding: Yuhan Dai (emilydai0329@gmail.com)

diagnosis [1], drug development [2], and biological experimental analysis. For example, in scenarios such as pathological slide analysis, cellular structure observation, and evaluation of drug efficacy, image data has become an important source of information for supporting medical research and clinical decision-making [3]. However, medical and biopharmaceutical images often exhibit characteristics such as complex structures, low contrast, uneven illumination, and large variations in the scale of abnormal regions. As a result, traditional image analysis methods that rely on manual experience or shallow features struggle to achieve stable and high-precision anomaly detection, which to some extent limits the development of intelligent healthcare and automated biopharmaceutical analysis [4]. In recent years, with the advancement of deep learning technologies, visual models based on convolutional neural networks and Transformers have achieved significant progress in the field of medical image analysis. Nevertheless, existing methods still face several key challenges when processing medical and biological images, such as the interference of illumination variations with structural information representation, insufficient utilization of frequency-domain information, and limited capability of models to capture global anomalous patterns. These issues can negatively affect the accuracy and robustness of anomaly detection [5].

To address the above challenges, researchers have gradually begun to explore image modeling methods that integrate multi-domain information, aiming to enhance the model's perception of complex anomalous patterns by combining spatial structural features with frequency features. However, in the task of medical and biopharmaceutical image anomaly detection, several critical problems remain to be solved, including how to effectively suppress the interference caused by illumination variations [6], how to mine stable and discriminative frequency features, and how to strengthen the representation of anomalous regions while maintaining the capability of global semantic understanding [7]. Therefore, constructing a unified anomaly detection framework that integrates structural information, frequency information, and global contextual relationships is of great significance for improving the level of intelligence in medical and biopharmaceutical image analysis [8].

Based on the above research background and challenges, this paper proposes a medical and biopharmaceutical image anomaly detection method that integrates Retinex state-space duality, a frequency-domain consensus-driven mechanism, and a Vision Transformer. First, the illumination and structural components of the image are separated and reconstructed through Retinex state-space dual modeling, thereby enhancing the structural representation of anomalous regions while reducing the impact of illumination variations. Subsequently, a frequency-domain consensus-driven mechanism is introduced to model the consistency relationships among features at different scales in the frequency space, so as to strengthen discriminative frequency information. On this basis, the self-attention mechanism of the Vision Transformer is utilized to model global features, enabling the capture of potential long-range dependencies and anomalous patterns in medical images. Through the collaborative modeling of multi-domain information, this study aims to improve the accuracy and robustness of anomaly detection in medical and biopharmaceutical images, providing an effective technical approach for automated image inspection in intelligent medical imaging analysis and biopharmaceutical research.

The main contributions of this paper are as follows.

(1) First, this study introduces a Retinex state-space duality mechanism to alleviate the challenges of anomaly detection caused by the coupling of uneven illumination and structural information in medical and biopharmaceutical images. Based on traditional Retinex decomposition, the illumination component and structural component are modeled in a dual manner within the state space, enabling the coordinated representation of brightness variations and tissue texture features. In this way, illumination interference can be suppressed while the structural information of anomalous regions is enhanced. This mechanism improves the structural representation capability of images during the feature extraction stage and provides more stable and discriminative feature representations for subsequent Transformer-based models, thereby improving the accuracy and robustness of anomaly detection in medical and biopharmaceutical images.

(2) Second, this paper proposes an FCD (Frequency Consensus-Driven) mechanism to enhance the model's perception of critical frequency features in medical images. The proposed approach models the consistency among features at different scales in the frequency domain, enabling the extraction of stable and discriminative frequency information from the image and strengthening the saliency of anomalous regions in frequency representations. At the same time, the frequency consensus mechanism guides the model to focus on consistent

frequency patterns during feature learning, reducing the interference of noise and irrelevant information. This process provides a more stable and effective frequency representation for the global feature modeling of the Transformer, thereby further improving the accuracy and robustness of anomaly detection in medical and biopharmaceutical images.

(3) Third, Vision Transformer (ViT) is incorporated into the model framework to enhance the global modeling capability for complex anomalous patterns in medical and biopharmaceutical images. Unlike traditional convolutional neural networks that mainly rely on local receptive fields, ViT captures long-range dependencies in images through the self-attention mechanism, enabling a more comprehensive understanding of overall structures and potential anomalous features. In this study, ViT is combined with the Retinex state-space duality mechanism and the frequency consensus-driven mechanism, allowing the model to jointly integrate spatial structural information and frequency feature representations. As a result, the proposed framework improves the recognition ability for subtle anomalous regions and further enhances the accuracy and robustness of anomaly detection in medical and biopharmaceutical images.

The logical structure of this paper is organized as follows.

In Section 2, we introduce the related work and provide an overview of previous studies in this field, including their advantages and limitations. In Section 3, we present the main methodology of this paper, including the Retinex state-space duality mechanism, the FCD (Frequency Consensus-Driven) frequency-domain consensus mechanism, and the Vision Transformer (ViT) module. In Section 4, we discuss the experimental results, including comparative experiments and ablation studies, and provide visualization analyses. Finally, in Section 5, we present the discussion and conclusions, where the limitations of the proposed method are analyzed, the main findings are summarized, and future research directions are outlined.

2. Related Work

In recent years, with the continuous development of medical imaging technology and biopharmaceutical research [9], medical and biopharmaceutical data have played an increasingly important role in disease diagnosis [10], pathological analysis, and drug development processes [11]. Through the automated analysis of microscopic images, pathological slice images, and various types of medical imaging data [12], it is possible to effectively improve the efficiency of disease identification and reduce the subjectivity associated with manual diagnosis [13]. Therefore, how to utilize data science and deep learning methods to accurately identify abnormal regions in medical and biopharmaceutical images has become an important research direction in the field of intelligent medical image analysis. Especially in imaging environments characterized by complex tissue structures, low contrast, and strong noise interference, anomaly detection technology is of great significance for improving the accuracy and reliability of medical image analysis.

In related studies, traditional anomaly detection methods for medical and pharmaceutical images mainly rely on manually designed features, such as texture features, morphological features, and statistical features, combined with machine learning models such as support vector machines and random forests for classification and recognition. These methods achieved certain success in early medical image analysis; however, due to the complex and diverse structures of medical and biological images, handcrafted features are often insufficient to fully represent the deep semantic information contained in images, which limits their detection performance to some extent [14]. With the development of deep learning technology, medical and pharmaceutical image analysis methods based on convolutional neural networks have gradually become mainstream. Through end-to-end feature learning, these approaches can effectively improve the accuracy of anomaly detection. For example, network architectures such as ResNet and U-Net have been widely applied in biopharmaceutical classification and segmentation tasks, achieving significant progress in tumor detection, cell recognition, and pathological image analysis [15]. However, convolutional neural networks mainly rely on local receptive fields during feature extraction, and their ability to model long-range dependencies in images is limited, which to some extent affects the model's overall understanding of complex anomalous structures.

To further improve the capability of biopharmaceutical image analysis, researchers have proposed various deep learning methods for tasks such as anomaly detection, image segmentation, and object recognition, aiming

to enhance the level of automation in medical image analysis. Dalmonte et al. proposed a Q-Former Autoencoder framework [16], which utilizes a pretrained visual foundation model as the feature extractor and employs a Q-Former structure as a bottleneck module to aggregate multi-scale features, while incorporating perceptual loss to improve reconstruction quality, thereby enabling unsupervised medical image anomaly detection. The advantage of this method lies in its ability to effectively leverage the high-level semantic representation capability of pretrained foundation models and achieve satisfactory anomaly detection performance without requiring large amounts of labeled data. However, the method still relies primarily on reconstruction error for anomaly identification, which limits its sensitivity to complex structural anomalies or subtle abnormal regions. In addition, it lacks sufficient modeling of frequency-domain information and illumination variations. To address the challenges of complex structures and multi-scale feature representation in medical images, Alrfou et al. proposed the GC-UNet network [17], which introduces a Global Context Vision Transformer into both the encoder and decoder to integrate global self-attention with local feature modeling, thereby improving the accuracy of medical image segmentation. The method achieved good segmentation performance on multiple medical datasets. However, the study mainly focuses on medical image segmentation tasks, and its capability to identify potential abnormal regions in anomaly detection scenarios remains limited. Moreover, the model primarily emphasizes spatial structural feature learning and lacks systematic modeling of frequency features and illumination variations. In terms of improving the efficiency of medical image processing, Martínez et al. approached the problem from the perspective of computational architecture and proposed a medical image processing method based on Processing-in-Memory (PIM) [18]. On a real PIM hardware platform, they implemented several fundamental algorithms, including convolution, threshold processing, and histogram computation, thereby significantly reducing the computational bottleneck caused by data movement. This study is of great significance for improving the efficiency of medical image processing. However, its primary focus lies in computational efficiency and hardware architecture optimization, and the enhancement of high-level semantic understanding and anomaly detection capability in medical images still depends on subsequent algorithmic models. Focusing on the problem of drug recognition in the biopharmaceutical field, Sachin et al. proposed a hybrid deep learning framework that combines YOLOv12n for pill detection and ConvNeXt-Tiny for multi-attribute classification [19], while introducing a metadata validation mechanism to achieve high-precision pill recognition and attribute prediction. This method demonstrates high accuracy and robustness in drug recognition tasks. However, its research mainly focuses on object detection and classification scenarios and pays limited attention to the identification of abnormal regions in complex medical images. In addition, its feature modeling primarily relies on spatial-domain information. Based on the development of object detection technology, Qomariah et al. further utilized the YOLOv8 model to perform automatic pill counting tasks and evaluated the model performance under different training epochs [20]. Experimental results show that YOLOv8 achieves high accuracy and good generalization capability in real-time detection tasks. Nevertheless, this method mainly targets the detection and counting of regular objects and still has limited capability in detecting abnormal regions with complex structures and blurred boundaries in medical and biological images.

Overall, although current anomaly detection methods for medical and biopharmaceutical images have made significant progress in deep learning and visual modeling, they still face limitations in suppressing illumination variation interference, modeling the consistency of frequency-domain features, and achieving collaborative representation of spatial and frequency information. Particularly in complex medical imaging environments, how to enhance the structural representation of anomalous regions while improving global modeling capability, and how to fully exploit frequency-domain information to strengthen feature discriminability, remain important research problems worthy of further investigation. Therefore, it is necessary to construct a unified framework that integrates Retinex-based structural enhancement, frequency-domain feature modeling, and the global representation capability of Transformers, in order to further improve the performance and robustness of anomaly detection in medical and biopharmaceutical images.

3. Method

To effectively improve the accuracy and robustness of anomaly detection in medical and biopharmaceutical images, this paper proposes a multi-domain feature modeling method that integrates Retinex state-space duality, an FCD (Frequency Consensus-Driven) mechanism, and a Vision Transformer. As shown in Figure 1. The proposed method first separates and enhances the structural and illumination information in images through Retinex state-space dual modeling, thereby reducing the influence of illumination variations on the recognition of anomalous regions. Subsequently, a frequency consensus-driven mechanism is introduced to explore stable consistency among multi-scale features in the frequency domain, thereby strengthening discriminative frequency features. On this basis, the self-attention mechanism of the Vision Transformer is utilized to model global features and capture long-range dependencies within images. Through the collaborative modeling of spatial-domain information, frequency-domain information, and global semantic representations, the proposed method can more effectively characterize potential anomalous patterns in medical and biopharmaceutical images, thereby improving the overall anomaly detection performance.

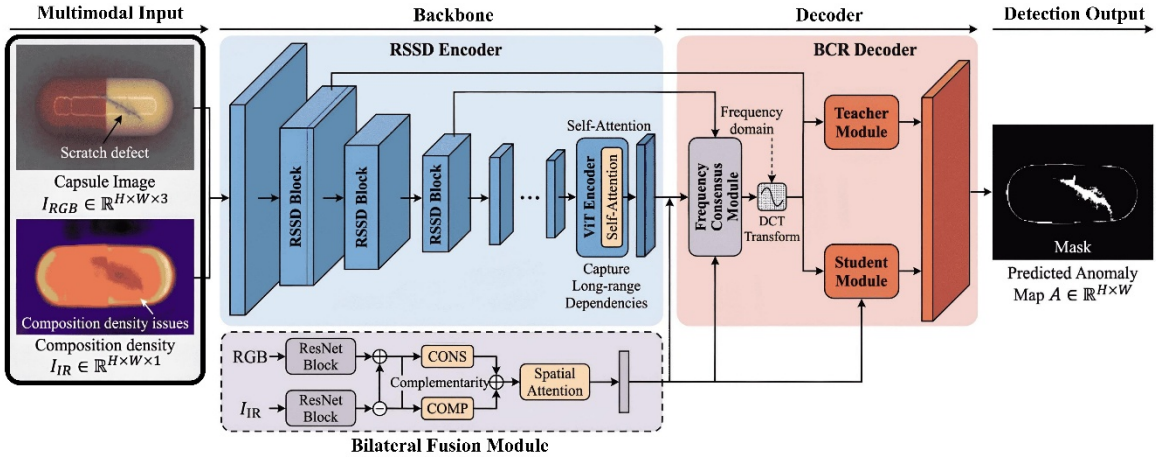

Figure 1. Overall algorithm architecture.

3.1. Retinex State-Space Duality

As shown in Figure 2, in the task of anomaly detection in medical and biopharmaceutical images, images are often affected by factors such as uneven illumination, complex tissue structures, and differences in imaging devices, which may cause the structural information of abnormal regions to be obscured by brightness variations. To address this issue, this paper proposes a Retinex state-space duality modeling method based on the Retinex theory. By collaboratively modeling the illumination component and structural component of the image within a state-space framework, the proposed approach enhances the structural representation of anomalous regions while suppressing illumination interference. First, according to the classical Retinex imaging model, the input image can be expressed as the multiplicative relationship between the reflectance component and the illumination component:

$$I(x,y)=R(x,y)\cdot L(x,y)$$

Here, $I(x,y)$ represents the observed intensity of the input medical image at the pixel location (x,y), $R(x,y)$ denotes the reflectance component that reflects the structural and texture information of tissues, and $L(x,y)$ represents the illumination component, describing the brightness variations in the imaging environment. To facilitate subsequent modeling, a logarithmic transformation is applied to this model to obtain an additive representation:

$$\log I(x,y)=\log R(x,y)+\log L(x,y)$$

Here, $\log R(x,y)$ represents the expression of structural information in the logarithmic space, while $\log L(x,y)$ denotes the logarithmic representation of illumination information. Through this transformation, the multiplicative coupling relationship can be converted into a linearly separable form, providing a foundation for state-space modeling. Furthermore, in order to describe the dual relationship between illumination and structure

within a dynamic feature space, this paper constructs a Retinex state-space model:

$$\mathbf{s}_t = \mathbf{A}\mathbf{s}_{t-1} + \mathbf{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{s}_t + \boldsymbol{\eta}_t$$

Here, $\mathbf{s}_t$ denotes the latent state vector at the t feature layer, which represents the joint state of structure and illumination; $\mathbf{A}$ is the state transition matrix that describes the evolution relationship of states across different layers; $\mathbf{u}_t$ represents the input feature vector; $\mathbf{B}$ is the input control matrix; $\mathbf{C}$ is the observation matrix; and $\mathbf{y}_t$ denotes the output feature representation. $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ represent the process noise and observation noise, respectively. Through the above state-space modeling, the variation relationship between image structure and illumination can be dynamically described in a multi-layer feature space. To achieve dual enhancement between the structural component and the illumination component, this paper further introduces a structure–illumination coupling constraint function:

$$\mathcal{L}_{\text{dual}} = \int_{\Omega} \left\| \nabla R(x,y) - \sqrt{\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \nabla^2 I(i,j)} \right\|^2 dxdy$$

Here, Ω denotes the image spatial region, $\nabla R(x,y)$ represents the gradient operator of the reflectance component used to extract structural edge information, $\nabla^2 I(i,j)$ denotes the Laplacian operator of the input image reflecting the local variation intensity of the image, and $|\Omega|$ represents the number of pixels within the region. This constraint enables the reflectance component to focus more on structural variations in the image, thereby enhancing the boundary representation of anomalous regions. On this basis, in order to further improve the stability of structural information, this paper designs a structural enhancement function:

$$R^*(x,y) = \frac{R(x,y)}{\sqrt{1+\lambda\sum_{k=1}^{K}\left|\nabla^k L(x,y)\right|^2}}$$

Here, $R^*(x,y)$ represents the enhanced structural component, λ is a regularization parameter used to control the degree of illumination suppression, K denotes the number of multi-order gradient scales, and $\nabla^k L(x,y)$ represents the gradient variation of the illumination component at the k scale. This formulation enables adaptive suppression of brightness disturbances in regions with significant illumination variations, thereby strengthening the representation of true structural information. Finally, in order to integrate the Retinex state-space dual representation with deep feature extraction, this paper defines the final feature mapping function:

$$\mathbf{F}_{\text{ret}} = \sum_{i=1}^{N} \alpha_i \phi_i\left(R^*(x,y)\right) + \beta_i \psi_i\left(L(x,y)\right)$$

Here, $\mathbf{F}_{\text{ret}}$ represents the Retinex dual feature representation, N denotes the number of feature layers, $\phi_i(\cdot)$ and $\psi_i(\cdot)$ represent the structural feature mapping function and the illumination feature mapping function, respectively, and α_i and β_i are learnable weight parameters used to balance the contributions of structural and illumination information. Through the above Retinex state-space dual modeling, the model is able to achieve the collaborative representation of structural information and illumination information at the feature level, thereby more effectively highlighting the structural characteristics of anomalous regions in complex medical and biopharmaceutical imaging environments.

3.2. Frequency Consensus-Driven

As shown in Figure 3, in medical and biopharmaceutical image anomaly detection tasks, abnormal regions are often accompanied by texture variations, structural mutations, and inconsistencies in local frequency distributions. Therefore, relying solely on spatial-domain features is insufficient to fully capture the underlying anomalous information. To address this issue, this paper proposes a Frequency Consensus-Driven (FCD) mechanism, which models the consistency relationships among multi-scale features in the frequency domain to enhance discriminative frequency representations. Specifically, for an input feature map $F(x,y)$, a two-dimensional Discrete Fourier Transform (DFT) is applied to project the spatial-domain features into the frequency domain, resulting in the spectral representation:

$$\hat{F}(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} F(x,y)\, e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}$$

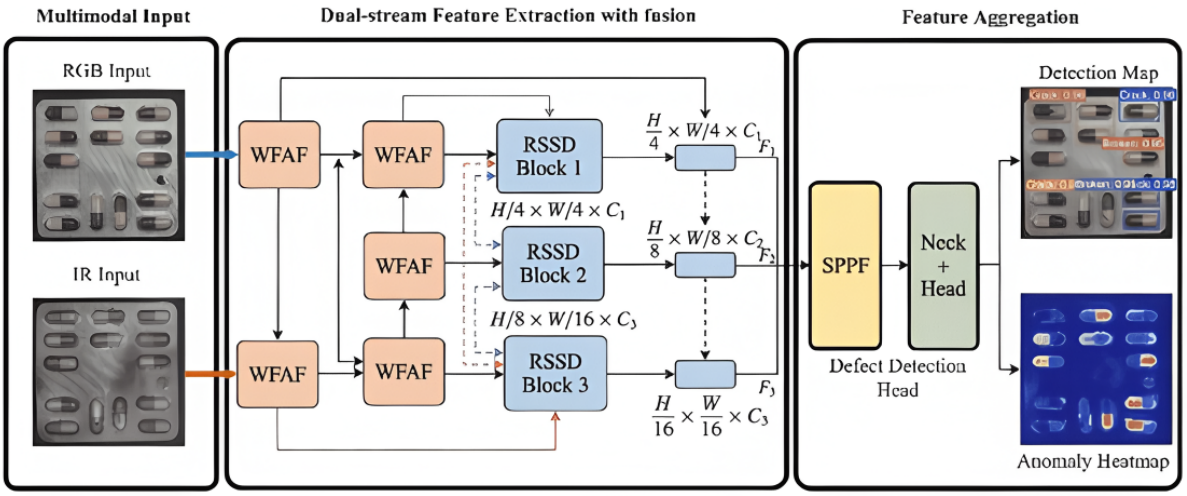

Figure 2. Diagram of Retinex model algorithm.

Here, $F(x,y)$ denotes the feature value of the spatial-domain feature map at position (x,y), and $\hat{F}(u,v)$ represents the complex spectral representation at the frequency coordinate (u,v) in the frequency domain. H and W denote the height and width of the feature map, respectively, and j is the imaginary unit. Through this transformation, the energy distribution characteristics of the image at different frequency scales can be obtained. To measure the importance of different frequency components, this paper defines a frequency energy response function:

$$E(u,v)=\sqrt{\left(\Re(\hat{F}(u,v))\right)^2+\left(\Im(\hat{F}(u,v))\right)^2}$$

Here, $E(u,v)$ denotes the magnitude energy at the frequency (u,v), while $\Re(\cdot)$ and $\Im(\cdot)$ represent the real and imaginary parts of a complex number, respectively. This formulation reflects the contribution of different frequency components to the structural representation of the image. To further explore the consensus relationships among multi-scale frequencies, this paper introduces a frequency-domain consensus function:

$$C(u,v)=\frac{1}{K}\sum_{k=1}^{K}\frac{E_k(u,v)}{\sqrt{\sum_{(i,j)\in\Omega}E_k(i,j)^2}}$$

Here, $C(u,v)$ denotes the consensus strength at the frequency location (u,v), K represents the number of multi-scale feature levels, $E_k(u,v)$ denotes the frequency energy response at the k scale, and Ω represents the entire frequency space region. This formulation performs consistency alignment of frequency energies across different scales through a normalization operation, thereby highlighting frequency patterns that exhibit stable responses across multiple scales. To further enhance the contribution of key frequency features to anomaly detection, this paper additionally designs a frequency-domain weight modulation function:

$$W(u,v)=\frac{\exp\left(\gamma C(u,v)\right)}{\sum_{(i,j)\in\Omega}\exp\left(\gamma C(i,j)\right)}$$

Here, $W(u,v)$ denotes the frequency weight coefficient, and γ is a temperature regulation parameter used to control the smoothness of the weight distribution. Through the Softmax normalization mechanism, this function assigns larger weights to frequency components with higher consensus strength, thereby enhancing their contribution to feature reconstruction. Finally, the frequency-domain consensus feature representation is obtained through a frequency-domain weighted reconstruction mechanism:

$$F_{\text{fcd}}(x,y)=\iint_{\Omega}W(u,v)\hat{F}(u,v)e^{j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}dudv$$

Here, $F_{\text{fcd}}(x,y)$ denotes the reconstructed spatial-domain feature representation obtained through the frequency consensus-driven mechanism, $W(u,v)$ is the frequency weighting function, and $\hat{F}(u,v)$ represents the spectral feature. This process is equivalent to performing a weighted inverse Fourier transform on the frequency-domain features, allowing frequency components with high consensus to dominate the reconstructed representation.

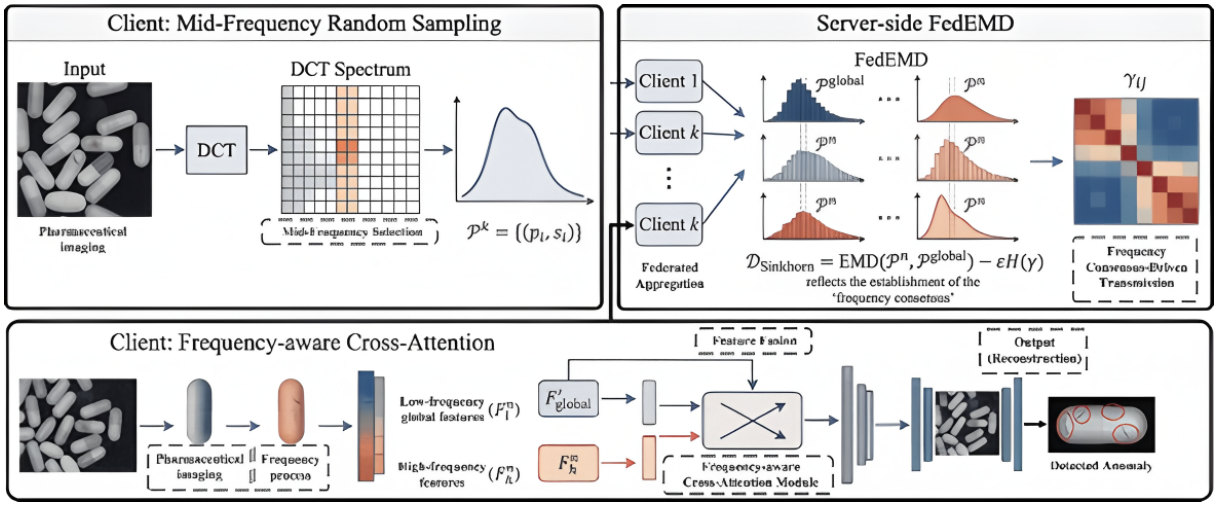

Figure 3. Diagram of FCD model algorithm.

3.3. Vision Transformer

As shown in Figure 4. In this study, a Vision Transformer (ViT) module is introduced into the overall framework. Through the self-attention mechanism, global modeling of image features is performed, thereby enhancing the model's ability to perceive potential anomalous regions. Specifically, for the input feature map $F \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels of the feature map, respectively, the feature map is first divided into N image patches of size $P\times P$. These patches are then linearly embedded to obtain a sequence representation.

$$z_0^i = \mathbf{E}\cdot \mathrm{vec}(F_i) + \mathbf{p}_i$$

Here, $F_i \in \mathbb{R}^{P\times P\times C}$ denotes the i image patch, vec(·) represents the vectorization operation, $\mathbf{E} \in \mathbb{R}^{(P^2C)\times D}$ is the learnable embedding matrix, D denotes the embedding dimension, and $\mathbf{p}_i$ is the positional encoding vector used to preserve the spatial position information of each patch in the original image. Through this process, the input sequence $Z_0 = \{z_0^1, z_0^2, \ldots, z_0^N\}$ can be obtained. In the Transformer encoder layer, global modeling of the sequence features is first performed through the Multi-Head Self-Attention (MHSA) mechanism, whose basic computation form is given as follows:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Here, $Q = ZW_Q$, $K = ZW_K$, and $V = ZW_V$ denote the query matrix, key matrix, and value matrix, respectively, where W_Q, W_K, and $W_V \in \mathbb{R}^{D\times d_k}$ are learnable parameter matrices, and d_k represents the dimension of the key vectors. The term $\sqrt{d_k}$ serves as a scaling factor to prevent excessively large inner-product values that may lead to gradient instability. To further enhance the model's representation capability, ViT adopts a multi-head attention mechanism, in which the outputs of multiple attention heads are concatenated:

$$\mathrm{MSA}(Z) = \mathrm{Concat}\left(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h\right)W_O$$

Here, $\mathrm{head}_i = \mathrm{Attention}(ZW_Q^i, ZW_K^i, ZW_V^i)$ denotes the output of the i attention head, h represents the number of attention heads, and W_O is the output projection matrix. Through the multi-head mechanism, correlations among image features can be captured from different subspaces. Subsequently, in each Transformer encoder layer, a Feed Forward Network (FFN) is employed to further enhance the feature representation capability:

$$\mathrm{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2$$

Here, x denotes the input feature vector, W_1 and W_2 are learnable weight matrices, b_1 and b_2 represent the bias terms, and $\sigma(\cdot)$ denotes the nonlinear activation function. This structure enhances the discriminative capability of the features through nonlinear mapping. To improve the stability of the model, ViT introduces residual connections and layer normalization in each encoder layer, and the update process can be expressed as follows:

$$Z_{l+1} = Z_l + \mathrm{MSA}\left(\mathrm{LayerNorm}(Z_l)\right) + \mathrm{FFN}\left(\mathrm{LayerNorm}(Z_l)\right)$$

Here, Z_l denotes the input feature representation of the l layer, and Z_{l+1} represents the updated feature representation. LayerNorm(·) denotes the layer normalization operation, which is used to stabilize the training process. By stacking multiple Transformer encoder layers, the model can progressively capture the global dependency relationships among image patches. Finally, the obtained global feature representation can be expressed as:

$$F_{\text{vit}} = \frac{1}{N}\sum_{i=1}^{N} Z_L^i$$

Here, Z_L^i denotes the feature representation of the i image patch in the output of the L Transformer layer, N represents the number of image patches, and F_{vit} is the final global feature vector. This global representation can effectively integrate long-range structural information in medical and biological images, thereby enhancing the model's capability to identify complex anomalous regions and providing richer and more discriminative feature representations for subsequent anomaly detection modules.

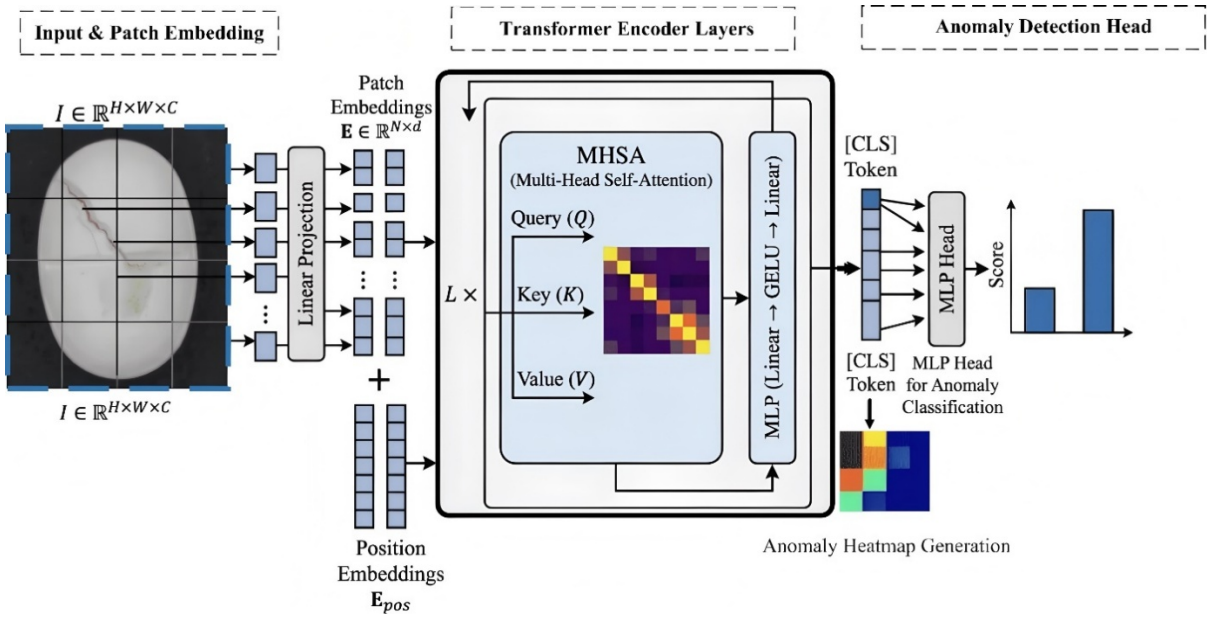

Figure 4. Diagram of Vision Transformer model algorithm.

4. Experiment

4.1. Experimental Environment

All experiments were conducted on a machine equipped with an Intel Xeon Gold 6230 CPU running at 2.10 GHz, 256 GB of RAM, and an NVIDIA Tesla V100 GPU with 32 GB of memory. The operating system used was Ubuntu 20.04. The models were implemented in Python using PyTorch (version 1.12.1), along with supporting libraries such as NumPy and SciPy for mathematical computations and data processing. The training and testing of the models leveraged the GPU for efficient matrix operations and parallelized data handling. Hyperparameter tuning was performed using a grid search, exploring different configurations for learning rates, batch sizes, and attention head numbers. The final models were trained using a batch size of 512, a learning rate of 0.001, and 8 attention heads over 12 transformer layers. The models were trained over 100 epochs, with early stopping applied if no improvement was observed in the validation loss after 10 consecutive epochs.

4.2. Experimental Data

- ePillID Dataset [21]

The ePillID dataset is a low-shot fine-grained benchmark designed for pill identification tasks. It contains images of various pharmaceutical pills with subtle visual differences in shape, color, imprint, and size. The dataset focuses on scenarios where only a limited number of labeled samples are available per class, making it suitable for evaluating few-shot and fine-grained recognition models. Due to the high similarity among pill

categories and the limited training samples, ePillID presents a challenging benchmark for developing robust pill identification and classification algorithms.

• C3PI RxIMAGE Dataset [22]

The C3PI RxIMAGE dataset is a pharmaceutical image dataset used for pill identification and medication-related visual analysis. It contains images of prescription pills collected from clinical and pharmaceutical information systems. The dataset includes diverse pill appearances with variations in color, texture, imprint, and lighting conditions. It is commonly used to evaluate computer vision methods for pill recognition, classification, and medical image analysis tasks, supporting the development of intelligent pharmaceutical information retrieval and verification systems.

• Pill Defect Dataset [23]

The Pill Defect dataset is designed for pharmaceutical quality inspection and defect detection tasks. It contains images of pills with both normal and defective samples, where defects may include cracks, chipping, surface contamination, deformation, or manufacturing inconsistencies. The dataset is often used to evaluate machine learning and deep learning methods for automated industrial inspection, particularly in pharmaceutical manufacturing environments where accurate and reliable quality control is essential.

• PillQC Dataset [24]

The PillQC dataset is a pharmaceutical pill quality control dataset developed for detecting defects in pill production processes. It includes images of pills with different types of anomalies such as scratches, breakages, irregular shapes, and surface imperfections. The dataset is intended to support research on automated defect detection using deep learning and anomaly detection methods, helping to improve the reliability and efficiency of pharmaceutical manufacturing quality assurance systems.

4.3. Evaluation Metrics

• Precision (P)

Precision measures the proportion of correctly predicted positive samples to all samples predicted as positive. In medical applications, this reflects the framework's ability to generate accurate diagnostic labels or relevant responses without including irrelevant or incorrect information. The formula is as follows:

$$P = \frac{TP}{TP + FP}$$

TP: True positives (correctly identified findings or answers).

FP: False positives (incorrectly identified findings or answers).

A high precision score is vital in medical scenarios where false positives can lead to unnecessary tests, treatments, or patient anxiety.

• Recall (R)

Recall quantifies the proportion of correctly predicted positive samples out of all actual positive samples. In medical contexts, it assesses the framework's ability to capture as many relevant findings as possible. The formula is as follows:

$$R = \frac{TP}{TP + FN}$$

FN: False negatives (missed findings or incorrect omissions)

Recall is especially critical in medical diagnostics, where missed diagnoses (false negatives) can have severe consequences for patient outcomes.

• F1-Score (F1)

F1-Score provides a harmonic mean of precision and recall, balancing these two aspects in scenarios where there is a trade-off. This metric is particularly relevant for evaluating medical tasks like diagnostic classification or report generation, where both precision (accuracy of findings) and recall (completeness of findings) are equally important.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

A high F1-score indicates that the framework can maintain a robust balance between precision and recall, making it well-suited for complex multimodal tasks.

- Accuracy (Acc)

Accuracy measures the proportion of correctly predicted samples out of the total number of samples. In the context of this study, accuracy serves as a baseline metric for assessing overall performance across various datasets.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

TN: True negatives (correctly rejected irrelevant findings).

4.4. Experimental Comparison and Analysis

From the experimental results presented in Table 1 and Figure 5, it can be observed that the proposed method achieves significantly better performance than other comparative methods on both medical and biopharmaceutical image datasets, ePillID and C3PI RxIMAGE, demonstrating the clear advantages of the proposed model in anomaly detection tasks. On the ePillID dataset, the proposed method achieves the best results across all four evaluation metrics. Specifically, the Accuracy reaches 94.27%, which represents an improvement of approximately 12.01 percentage points compared with the best-performing comparative method, X Liu et al., which achieves 82.26%. In terms of Precision, the proposed method attains 94.86%, improving by 11.75 percentage points over 83.11%. For the Recall metric, the proposed method reaches 93.52%, which is 12.05 percentage points higher than 81.47%. Regarding the comprehensive evaluation metric F1-score, the proposed method achieves 94.18%, representing an improvement of 11.90 percentage points compared with the second-best method, which achieves 82.28%. These results indicate that the proposed method has clear advantages in terms of accuracy and stability when identifying anomalous samples, enabling it to more effectively capture complex abnormal features in medical images. On the C3PI RxIMAGE dataset, the proposed method also demonstrates a significant performance advantage. The experimental results show that the proposed method achieves 92.64%, 93.31%, 91.87%, and 92.58% in Accuracy, Precision, Recall, and F1-score, respectively, all of which are the highest among the compared methods. Compared with the best-performing baseline method, X Liu et al. (with Accuracy of 84.62% and F1-score of 84.69%), the proposed method improves Accuracy by approximately 8.02 percentage points and F1-score by approximately 7.89 percentage points. Furthermore, compared with other methods such as Y Lee et al. (Accuracy 74.34%, F1-score 74.46%) and B Wang et al. (Accuracy 77.43%, F1-score 77.46%), the proposed method demonstrates a more pronounced overall performance advantage.

Table 1. Comparative analysis experiment demonstration for ePillID Dataset and C3PI RxIMAGE Dataset.

Method	Datasets							
	ePillID				C3PI RxIMAGE			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
YY Ou et al. [25]	62.48	63.15	61.27	62.2	65.81	66.42	64.39	65.39
Y Lee et al. [26]	78.12	79.04	77.42	78.22	74.34	75.12	73.81	74.46
Y Chen et al. [27]	68.53	67.59	69.91	68.73	71.07	70.15	72.24	71.18
X Liu et al. [28]	82.26	83.11	81.47	82.28	84.62	85.48	83.91	84.69
M Xu et al. [29]	71.64	72.32	70.86	71.58	68.29	69.04	67.62	68.32
BW et al. [30]	75.11	75.84	74.39	75.11	77.43	78.19	76.74	77.46
Ours	94.27	94.86	93.52	94.18	92.64	93.31	91.87	92.58

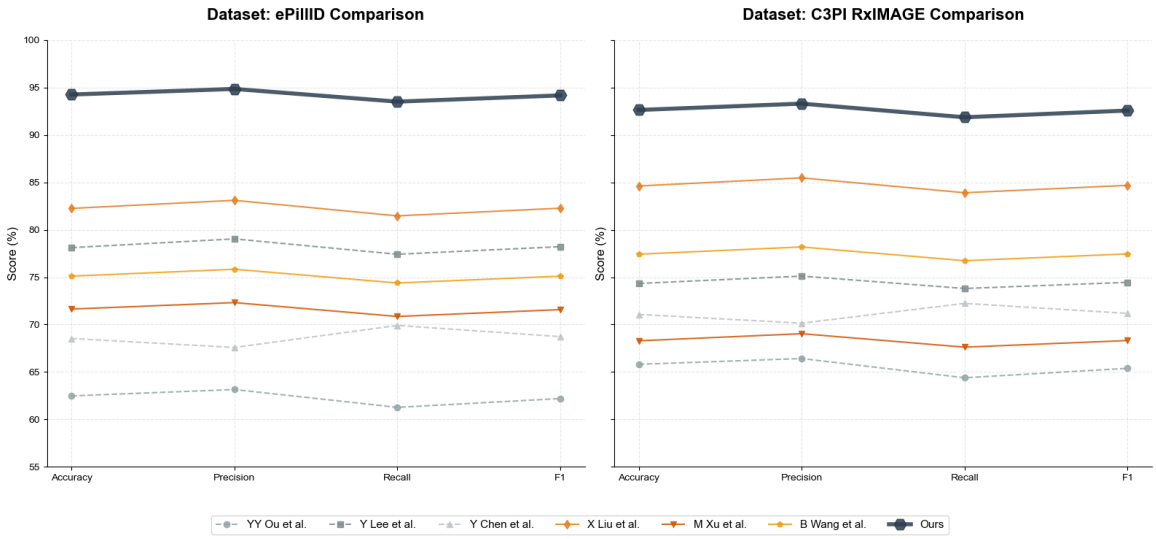

Figure 5. Visual comparison of experiments conducted on the ePillID Dataset and C3PI RxIMAGE Dataset [25–30].

From the experimental results presented in Table 2 and Figure 6, it can be observed that the proposed method achieves the best performance on both the Pill Defect and PillQC medical and biopharmaceutical image datasets, significantly outperforming existing methods across multiple evaluation metrics, which fully verifies the effectiveness and stability of the proposed approach in anomaly detection tasks [31]. On the Pill Defect dataset, the proposed method achieves the highest results across all four evaluation metrics. Specifically, the Accuracy reaches 94.13%, representing an improvement of 10.87 percentage points compared with the best-performing comparative method X Liu et al., which achieves 83.26% [32]. In terms of Precision, the proposed method attains 94.82%, improving by 10.67 percentage points over 84.15%. For the Recall metric, the proposed method reaches 93.26%, which is 10.69 percentage points higher than 82.57% [33]. Regarding the comprehensive evaluation metric F1-score, the proposed method achieves 94.03%, representing an improvement of 10.68 percentage points compared with the second-best method, which achieves 83.35%. In addition, compared with other methods such as Y Lee et al. (Accuracy 79.54%, F1-score 79.27%) and B Wang et al. (Accuracy 75.19%, F1-score 75.42%), the proposed method also demonstrates clear performance advantages. [34] These results indicate that the proposed model has stronger feature representation capability and higher recognition accuracy in pharmaceutical defect image anomaly detection tasks. On the PillQC dataset, the proposed method also demonstrates stable and significant advantages. The experimental results show that the proposed method achieves 92.57%, 93.19%, 91.64%, and 92.41% in Accuracy, Precision, Recall, and F1-score, respectively, all of which are the highest among the compared methods. Compared with the best-performing baseline method X Liu et al. (with Accuracy of 84.73% and F1-score of 84.85%), the proposed method improves Accuracy by approximately 7.84 percentage points and F1-score by approximately 7.56 percentage points. Furthermore, compared with other methods such as Y Lee et al. (Accuracy 76.15%, F1-score 76.15%) and B Wang et al. (Accuracy 78.61%, F1-score 78.63%), the proposed method also maintains a clear overall performance advantage.

Considering the experimental results across the four datasets, it can be observed that the proposed method achieves the best or significantly leading performance across all evaluation metrics and demonstrates strong stability and generalization ability across different datasets. This indicates that the proposed Retinex state-space duality, FCD frequency-domain consensus-driven mechanism, and Vision Transformer global modeling strategy effectively enhance the model's capability to represent structural features and frequency characteristics of anomalous regions in medical and biopharmaceutical images [35]. Meanwhile, the incorporation of global contextual modeling further improves the accuracy and robustness of anomaly detection, enabling more reliable and efficient anomaly detection performance in complex medical image environments [36].

Table 2. Comparative analysis experiment demonstration for Pill Defect Dataset and PillQC Dataset.

Method	Datasets							
	Pill Defect				Pil1QC			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
YY Ou et al. [25]	64.31	65.28	63.44	64.35	68.22	68.91	67.13	68.01
Y Lee et al. [26]	79.54	80.42	78.16	79.27	76.15	77.03	75.29	76.15
Y Chen et al. [27]	71.87	72.43	70.92	71.67	73.49	74.21	72.56	73.38
X Liu et al. [28]	83.26	84.15	82.57	83.35	84.73	85.62	84.09	84.85
M Xu et al. [29]	67.42	68.19	66.83	67.5	65.94	66.57	64.88	65.71
BW et al. [30]	75.19	76.34	74.52	75.42	78.61	79.45	77.82	78.63
Ours	94.13	94.82	93.26	94.03	92.57	93.19	91.64	92. 41

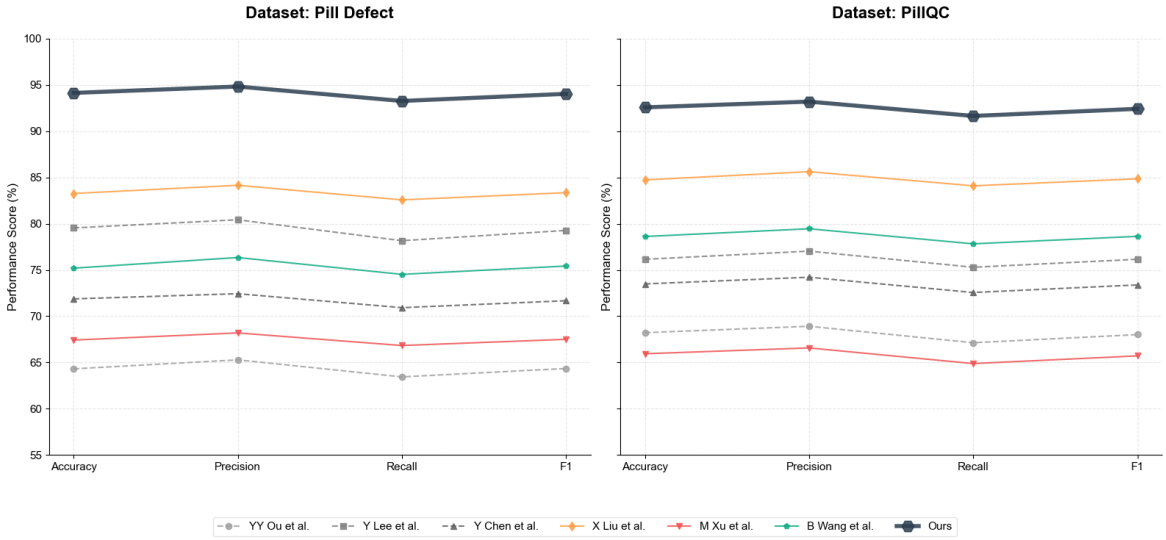

Figure 6. Visual comparison of experiments conducted on the Pill Defect Dataset and PillQC Dataset [25–30].

From the efficiency analysis results presented in Table 3 and Figure 7, it can be observed that the proposed method not only achieves superior detection performance on the ePillID and C3PI RxIMAGE datasets but also demonstrates clear advantages in computational efficiency and model complexity, highlighting the lightweight characteristics of the proposed approach [37]. This fully indicates the high efficiency and practical deployability of the proposed method in real-world medical and biopharmaceutical image anomaly detection tasks. On the ePillID dataset, the training time of the proposed method is 92.46 s, which is approximately 36.45 s shorter than Y Chen et al. (128.91 s) and 72.06 s shorter than X Liu et al. (164.52 s). Compared with Y Lee et al., which has the highest computational overhead (185.67 s), the training time is reduced by 93.21 s, demonstrating significantly higher training efficiency [38]. In terms of inference time, the proposed method requires only 104.38 ms, which is considerably lower than YY Ou et al. (168.12 ms), Y Lee et al. (192.45 ms), and B Wang et al. (186.21 ms). Compared with the relatively competitive method Y Chen et al. (154.37 ms), the inference time is further reduced by approximately 49.99 ms, indicating that the model can achieve faster real-time detection in practical deployment scenarios. In addition, regarding computational complexity, the proposed method requires only 12.17 G FLOPs, which is significantly lower than other methods, such as Y Chen et al. (20.84 G), X Liu et al. (25.32 G), and Y Lee et al. (28.16 G), representing a reduction in computational cost of nearly 40%–60%. In terms of model parameter size, the proposed method contains only 136.52 M parameters, which is significantly smaller than Y Lee et al. (276.41 M), B Wang et al. (265.82 M), and X Liu et al. (242.79 M). Even compared with the closest model in parameter size, Y Chen et al. (184.26 M), the

proposed method reduces the parameter scale by 47.74 M. Similar efficiency advantages are also observed on the C3PI RxIMAGE dataset. The training time of the proposed method is 95.83 s, which is significantly lower than Y Chen et al. (134.57 s), X Liu et al. (172.18 s), and Y Lee et al. (196.84 s). In terms of inference time, the proposed method requires only 108.72 ms, demonstrating a substantial speed advantage compared with 159.82 ms (Y Chen et al.) and 198.27 ms (Y Lee et al.). Meanwhile, regarding computational complexity, the proposed method requires 13.41 G FLOPs, which is significantly lower than those of other methods, such as 21.36 G, 26.84 G, and 29.53 G. In terms of parameter scale, the proposed method contains only 141.26 M parameters, which is considerably smaller than 254.31 M (X Liu et al.) and 288.17 M (Y Lee et al.), further demonstrating the efficiency and lightweight design of the proposed model [39].

Table 3. Comparative demonstration of efficiency analysis experiments for ePillID Dataset and C3PI RxIMAGE Dataset.

Method	Datasets							
	ePillID				C3PI RxIMAGE			
	Training Time (s)	Inference Time (ms)	Flops (G)	Para. (M)	Training Time (s)	Inference Time (ms)	Flops (G)	Para. (M)
YY Ou et al. [25]	142.34	168.12	22.47	215.38	155.62	174.53	24.19	228.64
Y Lee et al. [26]	185.67	192.45	28.16	276.41	196.84	198.27	29.53	288.17
Y Chen et al. [27]	128.91	154.37	20.84	184.26	134.57	159.82	21.36	191.53
X Liu et al. [28]	164.52	179.84	25.32	242.79	172.18	185.46	26.84	254.31
M Xu et al. [29]	151.28	163.59	23.61	221.54	162.34	167.91	24.72	230.42
BW et al. [30]	176.43	186.21	26.94	265.82	188.75	191.04	27.48	271.69
Ours	92.46	104.38	12. 17	136.52	95.83	108.72	13. 41	141. 26

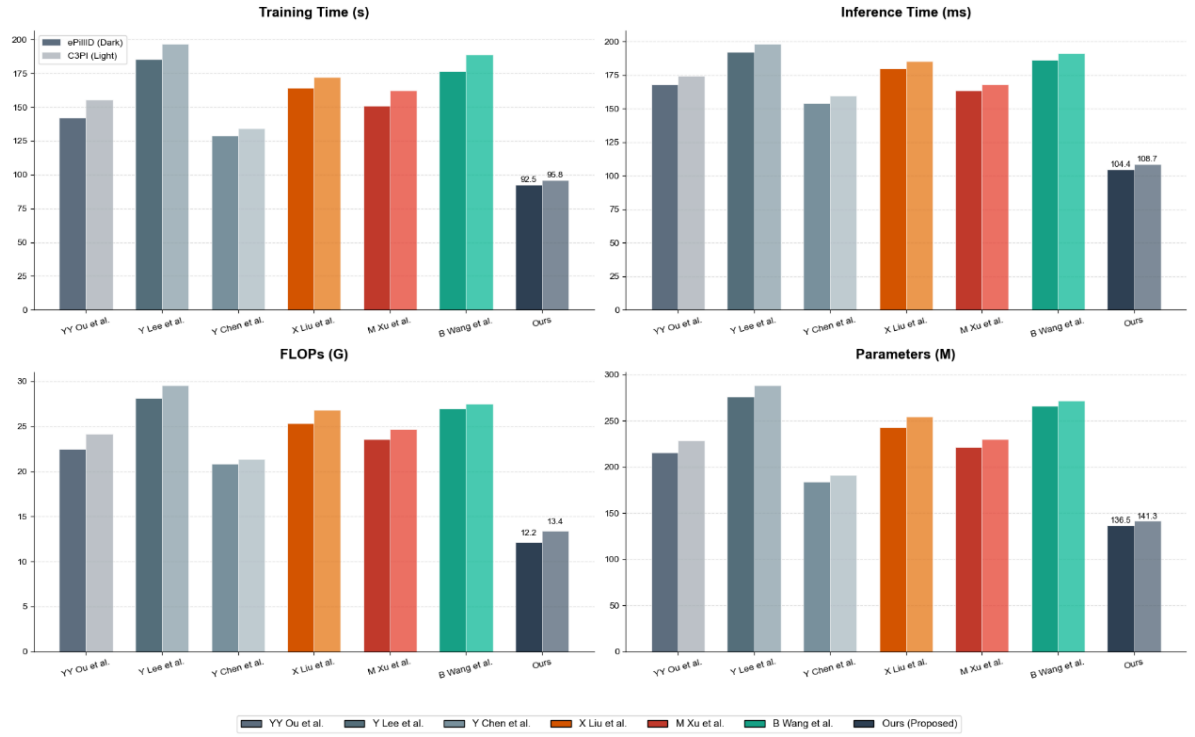

Figure 7. Visual comparison of efficiency analysis experiments for the ePillID Dataset and C3PI RxIMAGE Dataset [25–30].

From the efficiency analysis results presented in Table 4 and Figure 8, it can be observed that the proposed method also demonstrates clear computational efficiency advantages on the Pill Defect and PillQC datasets [40]. The proposed method outperforms existing comparative methods across multiple indicators, including training time, inference speed, computational complexity, and model parameter scale, further verifying that the proposed model maintains strong detection performance while also exhibiting favorable lightweight characteristics and practical deployment value. On the Pill Defect dataset, the training time of the proposed method is 92.34 s, which is significantly lower than that of other methods. For example, Y Chen et al. requires 125.48 s, M Xu et al. requires 148.91 s, X Liu et al. requires 162.37 s, B Wang et al. requires 174.56 s, and Y Lee et al. reaches 187.64 s. Compared with the relatively efficient method Y Chen et al., the proposed method still reduces the training time by 33.14 s, and compared with Y Lee et al., which has the largest training overhead, the training time is reduced by 95.30 s, demonstrating substantially higher training efficiency. In terms of inference time, the proposed method requires only 105.17 ms, which is significantly lower than Y Chen et al. (156.93 ms), X Liu et al. (178.41 ms), B Wang et al. (184.62 ms), and Y Lee et al. (194.27 ms). Compared with the relatively competitive method Y Chen et al., the inference speed is improved by approximately 51.76 ms, indicating that the model can achieve faster real-time inference in practical detection scenarios. Furthermore, regarding computational complexity (FLOPs), the proposed method requires only 11.42 G, which is substantially lower than other methods, such as 20.27 G (Y Chen et al.), 25.19 G (X Liu et al.), and 28.31 G (Y Lee et al.), representing a reduction in computational cost of approximately 40%–60%. In terms of model parameter scale, the proposed method contains only 135.86 M parameters, which is significantly smaller than Y Lee et al. (274.52 M), B Wang et al. (262.14 M), and X Liu et al. (243.85 M). Even compared with the relatively lightweight model Y Chen et al. (182.16 M), the proposed method still reduces the parameter size by 46.30 M. On the PillQC dataset, the proposed method maintains a consistent efficiency advantage. The training time is 96.71 s, which is significantly lower than 132.54 s (Y Chen et al.), 175.62 s (X Liu et al.), and 196.81 s (Y Lee et al.). In terms of inference time, the proposed method requires only 109.43 ms, whereas other methods generally range between 161.47 ms and 198.34 ms, indicating that the proposed method achieves higher real-time performance during the inference stage. Regarding computational complexity, the proposed method requires 12.87 G FLOPs, which is significantly lower than 21.84 G (Y Chen et al.), 26.73 G (X Liu et al.), and 29.57 G (Y Lee et al.). In terms of model parameter scale, the proposed method contains only 142.24 M parameters, whereas other methods generally range from 188.73 M to 286.19 M.

Table 4. Comparative demonstration of efficiency analysis experiments for Pill Defect Dataset and PillQC Dataset.

Method	Datasets							
	Pill Defect				Pil1QC			
	Training Time (s)	Inference Time (ms)	Flops (G)	Para. (M)	Training Time (s)	Inference Time (ms)	Flops (G)	Para. (M)
YY Ou et al. [25]	134.12	162.58	21.46	192.34	145.27	168.91	23.15	204.62
Y Lee et al. [26]	187.64	194.27	28.31	274.52	196.81	198.34	29.57	286.19
Y Chen et al. [27]	125.48	156.93	20.27	182.16	132.54	161.47	21.84	188.73
X Liu et al. [28]	162.37	178.41	25.19	243.85	175.62	185.06	26.73	251.94
M Xu et al. [29]	148.91	165.24	23.48	218.67	158.43	172.18	24.62	227.35
BW et al. [30]	174.56	184.62	26.85	262.14	186.29	190.73	27.91	273.46
Ours	92.34	105. 17	11.42	135.86	96.71	109.43	12.87	142.24

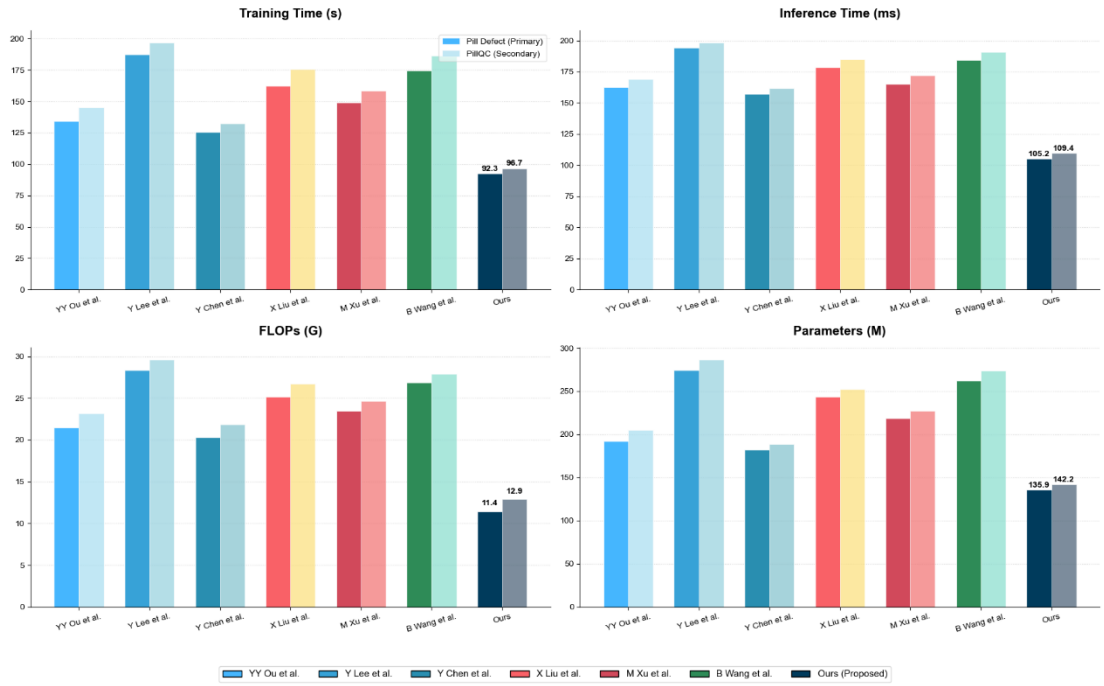

Figure 8. Visual comparison of efficiency analysis experiments for the Pill Defect Dataset and PillQC Dataset [25–30].

Considering the efficiency analysis results across all four datasets, it can be observed that the proposed method significantly reduces computational complexity and model size while maintaining high detection accuracy, and it substantially shortens both training and inference time. These results indicate that the proposed Retinex state-space duality modeling, frequency-domain consensus-driven mechanism, and Vision Transformer-based feature modeling strategy effectively improve feature representation efficiency while reducing redundant computation. As a result, the proposed approach achieves a strong balance between high accuracy and high efficiency in medical and biopharmaceutical image anomaly detection tasks, demonstrating strong potential for practical applications.

From the ablation study results presented in Table 5 and Figure 9, it can be observed that the three key modules proposed in this work—the Vision Transformer (ViT) architecture, Retinex state-space dual modeling, and the FCD (Frequency Consensus-Driven) mechanism—all play important roles in improving the overall performance. These modules work collaboratively and jointly contribute to the performance enhancement of the model in medical and biopharmaceutical image anomaly detection tasks. On the ePillID dataset, when the ViT module is removed, the model performance drops significantly, with Accuracy reaching only 64.27%, Precision 65.13%, Recall 63.48%, and F1-score 64.29%. Compared with the full model, which achieves Accuracy of 94.27% and F1-score of 94.18%, the performance decreases by 30.00 and 29.89 percentage points, respectively. This indicates that global contextual modeling plays a critical role in extracting anomalous features from pharmaceutical images. When the Retinex module is removed, the model performance decreases but still maintains a certain level, with Accuracy of 82.56% and F1-score of 82.53%, representing decreases of 11.71 and 11.65 percentage points compared with the full model. This result demonstrates that the Retinex state-space dual modeling effectively enhances the separation of image structure and illumination information, thereby improving the representation capability of anomalous regions. When the FCD frequency consensus-driven module is removed, the model performance further decreases to Accuracy of 73.18% and F1-score of 73.24%, representing reductions of 21.09 and 20.94 percentage points compared with the complete model, indicating that frequency-domain consistency constraints play an important role in stabilizing feature representation and enhancing anomaly detection capability. A similar trend is also observed on the C3PI RxIMAGE dataset. When the ViT module is removed, the model achieves Accuracy of 62.81% and F1-score of 62.72%, whereas the full

model reaches Accuracy of 92.64% and F1-score of 92.58%, indicating a performance drop of nearly 30 percentage points, which highlights the critical role of Transformer-based global dependency modeling in understanding complex pharmaceutical image structures. After removing the Retinex module, the model achieves Accuracy of 78.34% and F1-score of 78.28%, representing decreases of 14.30 percentage points in both metrics compared with the full model. When the FCD module is removed, the model obtains Accuracy of 75.62% and F1-score of 75.70%, approximately 17 percentage points lower than the complete model, indicating that frequency-domain information makes an important contribution to modeling complex textures and structural features. On the Pill Defect dataset, the complete model achieves the best results with Accuracy of 94.13% and F1-score of 94.03%, whereas the w/o ViT version reaches only Accuracy of 66.42% and F1-score of 66.26%, representing a performance drop of more than 27 percentage points. The w/o Retinex version achieves Accuracy of 80.17% and F1-score of 80.17%, decreasing by approximately 14 percentage points, while the w/o FCD version obtains Accuracy of 69.45% and F1-score of 69.53%, representing a decrease of approximately 24 percentage points. On the PillQC dataset, a consistent trend is also observed. The complete model achieves Accuracy of 92.57% and F1-score of 92.41%, whereas the w/o ViT version reaches only Accuracy of 61.35% and F1-score of 61.10%. The w/o Retinex version achieves Accuracy of 84.62% and F1-score of 84.43%, while the w/o FCD version obtains Accuracy of 71.94% and F1-score of 71.98%.

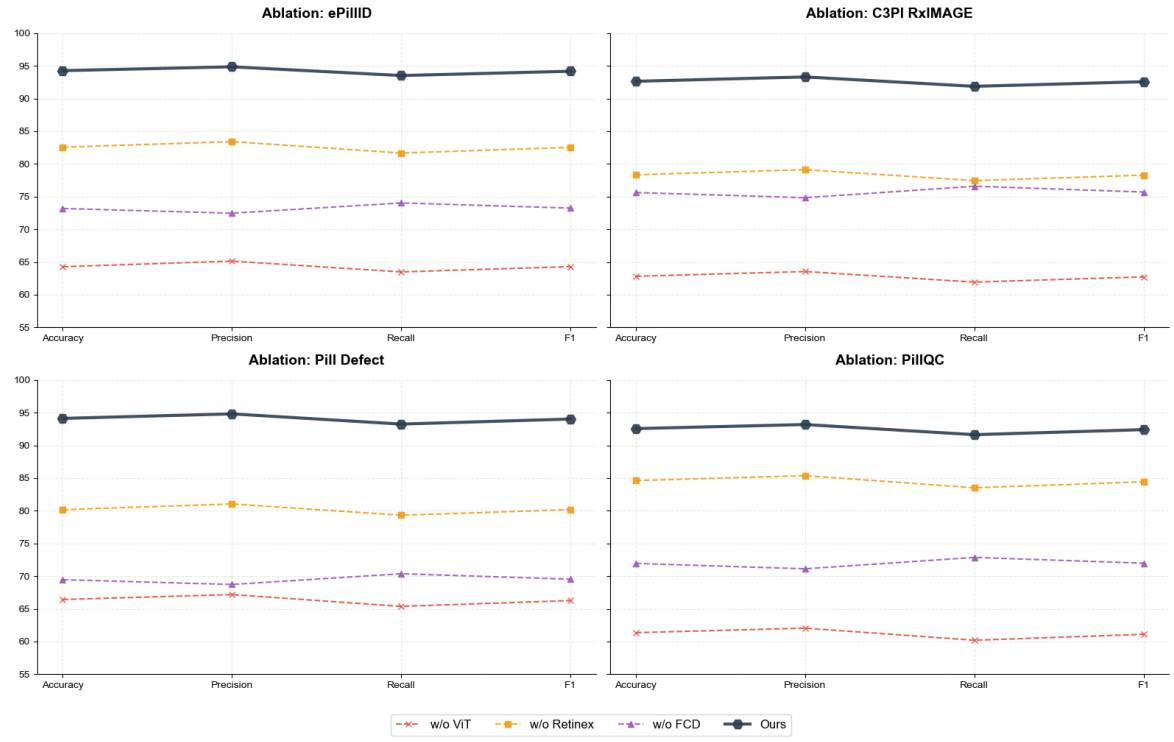

Figure 9. Visualization of ablation experiments on four datasets.

Considering the ablation study results across all four datasets, it can be observed that removing any key module leads to a noticeable performance degradation. Among them, the ViT module contributes the most to global semantic modeling, the Retinex state-space dual modeling effectively enhances the separation of image structure and illumination information, and the FCD frequency consensus-driven mechanism improves the stability and robustness of feature representation through frequency-domain consistency constraints. The collaboration of these three modules enables the complete model to achieve Accuracy and F1-score above 90% on all datasets, significantly outperforming all ablated versions, thereby fully demonstrating the effectiveness and rationality of the overall framework design proposed in this work.

Table 5. Results of Ablation Studies on Four Datasets.

Module	Datasets							
	ePillID				C3PI RxIMAGE			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
w/o ViT	64.27	65.13	63.48	64.29	62.81	63.54	61.92	62.72
w/o Retinex	82.56	83.41	81.67	82.53	78.34	79.12	77.45	78.28
w/o FCD	73.18	72.46	74.03	73.24	75.62	74.83	76.59	75.7
Ours	94.27	94.86	93.52	94.18	92.64	93.31	91.87	92.58
	Pill Defect				PillQC			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
w/o ViT	66.42	67.18	65.37	66.26	61.35	62.04	60.19	61.1
w/o Retinex	80.17	81.04	79.32	80.17	84.62	85.37	83.51	84.43
w/o FCD	69.45	68.72	70.36	69.53	71.94	71.12	72.86	71.98
Ours	94.13	94.82	93.26	94.03	92.57	93.19	91.64	92. 41

5. Discussion and Conclusions

This paper focuses on the task of anomaly detection in medical and biopharmaceutical images and proposes an anomaly detection framework that integrates Retinex state-space duality, a frequency consensus-driven (FCD) mechanism, and a Vision Transformer. Through the collaborative effects of spatial structure enhancement, frequency feature modeling, and global contextual relationship learning, the proposed framework improves the model's ability to recognize anomalies in complex medical imaging environments. The experimental results demonstrate that the proposed method achieves superior performance compared with several mainstream deep learning approaches across multiple medical image datasets, indicating that the method can effectively enhance the performance of anomaly detection in medical and biopharmaceutical images. Although the proposed method achieves promising results in the experiments, several limitations still exist. First, the Retinex state-space dual modeling and the frequency consensus-driven mechanism increase the computational complexity of the model to some extent. In scenarios involving high-resolution medical images or large-scale datasets, this may introduce additional computational overhead, potentially affecting the real-time performance of the model. Second, this study mainly validates the proposed method on publicly available medical image datasets, whereas medical images in real clinical environments often involve more complex imaging conditions and a wider variety of anomaly types. Therefore, the generalization capability of the model across different devices, imaging modalities, and cross-institutional datasets still requires further investigation. In addition, the proposed method primarily focuses on image-level anomaly detection and does not yet consider joint modeling of medical images with other modalities such as clinical information or genomic data. This limitation may restrict the potential application of the model in comprehensive medical analysis tasks.

Therefore, future research is needed to explore more efficient anomaly detection methods with stronger generalization capabilities. Further improvements can be pursued from several perspectives. On the one hand, lightweight network architectures or model compression techniques can be introduced to reduce computational complexity, enabling the model to better adapt to real-time medical detection scenarios. On the other hand, multimodal medical data fusion methods can be explored to jointly model medical images, clinical text, and bioinformatics data, thereby improving the model's ability to comprehensively understand complex disease characteristics. Furthermore, self-supervised learning or pretraining with foundation models could also be incorporated to further enhance the model's generalization ability and practical application value across different medical scenarios.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data analyzed in this study were obtained from two publicly available datasets: the ePillID Dataset and the C3PI RxIMAGE Dataset, as cited in the text.

Conflicts of Interest

The author declares no conflict of interest.

References

1 Shome D, Sarkar P, Etemad A. Region-Disentangled Diffusion Model for High-Fidelity PPG-to-ECG Translation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 15009–15019.

2 Yalcouyé A, Esoh K, Wonkam A. Enhancing Large-Scale Pharmacogenetic Studies in African Populations for Clinical Care and Drug Development. *Annual Review of Pharmacology and Toxicology* 2026; **66**: 171–190.

3 Lee J, Sulahria A, Luce S, *et al*. The Clinical Research Assistant. In *Clinical Research in Private Practice*; Academic Press: Cambridge, MA, USA, 2026; pp. 117–126.

4 Zhang Y, Mohsin SM, Mujlid H, *et al*. AI-Driven Blockchain Technology in Smart Healthcare System: Opportunities, Challenges and Future Implications. *Computer Science Review* 2026; **60**: 100909.

5 Xu D, Chen Y, Chai Z, *et al*. Knowledge Fusion in Deep Learning-Based Medical Vision-Language Models: A Review. *Information Fusion* 2026; **125**: 103455.

6 Zhang L, Li Z, Cheng L, *et al*. DLIENet: A Lightweight Low-Light Image Enhancement Network via Knowledge Distillation. *Pattern Recognition* 2026; **169**: 111777.

7 Xu M, Wei H, Nie Z, *et al*. Hybrid Dual-Heterogeneous Knowledge Distillation Network for Anomaly Detection in Retinal OCT Images. *IEEE Journal of Biomedical and Health Informatics* 2026. https://doi.org/10.1109/JBHI.2025.3650072

8 Verma RK. AI-Driven Image Enhancement Techniques for Edge Devices Balancing Quality and Performance. In *Advances in Image Processing*, Reliability, and Artificial Intelligence; Elsevier: Amsterdam, The Netherlands, 2026; pp. 151–168.

9 Irede EL, Aworinde OR, Lekan OK, *et al*. Medical Imaging: A Critical Review on X-ray Imaging for the Detection of Infection. *Biomedical Materials & Devices* 2026; **4**(**1**): 1–45.

10 Ahn S, Cho TS, Park SH. The Effects of Organizational Status on Market Entry Timing and Performance: The Biopharmaceutical Industry Amid Environmental Disruption. *Journal of Business Research* 2026; **205**: 115893.

11 Hui RW, Wu TK, Ho KC, *et al*. Large-Scale Profile Study on Hepatitis B Surface Antigen Levels in Chronic Hepatitis B: Implications for Drug Development Targeting Functional Cure. *Gut* 2026; **75**(**1**): 119–130.

12 Cakmak Y, Pacal I. A Comparative Analysis of Transformer Architectures for Automated Lung Cancer Detection in CT Images. *Journal of Intelligent Decision Making and Information Science* 2026; **3**: 528–539.

13 Sharma K, Hansen J, Susztak K, *et al*. Spatial Metabolomics and Multiomics Integration for Breakthroughs in Precision Medicine for Kidney Disease. *Nature Reviews Nephrology* 2026; **22**(**2**): 152–164.

14 Liu Y, Deng H, Fu J. DCM-Net: A Novel Dual-Branch CNN–Mamba Cross-Layer Feature Fusion Network for Medical Image Segmentation. *Biomedical Signal Processing and Control* 2026; **114**: 109267.

15 Yang J, Qiu P, Zhang Y, *et al*. D-Net: Dynamic Large Kernel with Dynamic Feature Fusion for Volumetric Medical Image Segmentation. *Biomedical Signal Processing and Control* 2026; **113**: 108837.

16 Dalmonte F, Bayar E, Akbas E, *et al*. Q-Former Autoencoder: A Modern Framework for Medical Anomaly Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2026, Tucson, AZ, USA, 6–10 March 2026.

17 Alrfou K, Zhao T. GC-UNet: Efficient Network for Medical Image Segmentation. *Multimedia Tools and Applications* 2026; **85**(**2**): 137.

18 Martínez H, Gómez-Luna J, Palomar R, *et al*. In-Memory Operators for Medical Image Processing. *Future Generation Computer Systems* 2026; **174**: 107939.

19 Sachin AS, Karthikeyan J. A Hybrid-Based Deep Learning Framework for Pill Detection, Multi-Attribute Classification and Prediction with Metadata Retrieval. *IEEE Access* 2026; **14**: 17905–17919.

20 Qomariah DU, Elvira AI, Kurniasari AA, *et al*. Automatic Pill Counting Using YOLOv8 to Improve Medication Distribution Accuracy. *International Journal of Public Health and Epidemiology* 2026; **5**(**2**): 28–33.

21 Usuyama N, Delgado NL, Hall AK, *et al*. ePillID Dataset: A Low-Shot Fine-Grained Benchmark for Pill Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2020, Seattle, WA, USA, 14–19 June 2020.

22 Peters L, Mortensen J, Nguyen T, *et al*. Enabling Complex Queries to Drug Information Sources through Functional Composition. *Studies in Health Technology and Informatics* 2013; **192**: 692.

23 Kim S, Jo Y, Cho J, *et al*. Spatially Variant Convolutional Autoencoder Based on Patch Division for Pill Defect Detection. *IEEE Access* 2020; **8**: 216781–216792.

24 Roy Choudhury K, Singh A, Padmini S. Detection of Pharmaceutical Pill Defects Through Deep One-Class Classification. In *International Conference on Data Analytics & Management*; Springer Nature: Singapore, 2024.

25 Ou YY, Tsai AC, Zhou XP, *et al*. Automatic Drug Pills Detection Based on Enhanced Feature Pyramid Network and Convolution Neural Networks. *IET Computer Vision* 2020; **14**(**1**): 9–17.

26 Lee Y, Lim H, Jang S, *et al*. Uniformaly: Towards Task-Agnostic Unified Framework for Visual Anomaly Detection. *Pattern Recognition* 2026; **169**: 111820.

27 Chen Y, Tao X, Chen B, *et al*. MPFR: Memory Prompt Feature Reconstruction for Continual Anomaly Detection and Segmentation. *Pattern Recognition* 2026; **175**: 112946.

28 Liu X, Wu C, Zhang H, *et al*. A Memory-Tree Driven Network for Multi-View Fusion Anomaly Detection. *Pattern Recognition* 2026; **170**: 112106.

29 Wang B, Wan J, Zhao J, *et al*. A Twin-Branch Decoupled Network for Multi-Class Unsupervised Anomaly Detection. *Engineering Applications of Artificial Intelligence* 2026; **167**: 113891.

30 Deng X, Oda S, Kawano Y. Graphene-Based Midinfrared Photodetector with Bull's Eye Plasmonic Antenna. *Optical Engineering* 2023; **62**(**9**): 097102.

31 Li J, Culver TB, Burgis CR, *et al*. Validating Nitrogen Removal Models with Field Bioretention Data. *Journal of Environmental Engineering* 2024; **150**(**8**): 04024037.

32 Yan H. Real-Time 3D Model Reconstruction Through Energy-Efficient Edge Computing. *Optimizations in Applied Machine Learning* 2022; **2**(**1**). https://doi.org/10.71070/oaml.v2i1.48

33 Li J, Culver TB, Persaud PP, *et al*. Developing Nitrogen Removal Models for Stormwater Bioretention Systems. *Water Research* 2023; **243**: 120381.

34 Yan H, Shao D. Multimodal Medical Image Analysis: Integrating LLM and RAG Deep Learning Strategies. *Journal of Advances in Information Technology* 2025; **16**(**4**): 568–581. https://doi.org/10.12720/jait.16.4.568-581.

35 Lu Y, Shao D, Ni X, *et al*. Emotion-Style Dual Prediction: A Multi-Task Deep Learning Approach for Artistic Images. *Cluster Computing* 2026; **29**(**1**): 31.

36 Deng X, Kawano Y. Surface Plasmon Polariton Graphene Midinfrared Photodetector with Multifrequency Resonance. *Journal of Nanophotonics* 2018; **12**(**2**): 026017.

37 Li J. Nitrogen Removal Models for Stormwater Bioretention Systems. *Ph.D. Thesis*, University of Virginia, Charlottesville, VA, USA, 2023.

38 Yan H, Shao D. Enhancing Transformer Training Efficiency with Dynamic Dropout. *arXiv* 2024, arXiv:

2411.03236. https://doi.org/10.48550/arXiv.2411.03236.

39 Luo Z, Yan H, Pan X. Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques. *Journal of Computational Methods in Engineering Applications* 2023; **3**: 1–12. https://doi.org/10.62836/jcmea.v3i1.030107.

40 Li J, Culver TB. Review of Process-Based Nitrogen Model for Agricultural Fields with Implications for Nitrogen Simulations in Stormwater BMPs. *Environmental Modelling & Software* 2022; **151**: 105363.