This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

USCNet: Transformer-Based Multimodal Fusion with Segmentation Guidance for Urolithiasis Classification

Changmiao Wang, Songqi Zhang, Yongquan Zhang, Yifei Wang, Liya Liu, Nannan Li, Xingzhi Li, Jiexin Pan, Yi Jiang, Xiang Wan, Hai Wang, Ahmed Elazab

***Abstract*— Kidney stone disease ranks among the most prevalent conditions in urology, and understanding the composition of these stones is essential for creating personalized treatment plans and preventing recurrence. Current methods for analyzing kidney stones depend on postoperative specimens, which prevents rapid classification before surgery. To overcome this limitation, we introduce a new approach called the Urinary Stone Segmentation and Classification Network (USCNet). This innovative method allows for precise preoperative classification of kidney stones by integrating Computed Tomography (CT) images with clinical data from Electronic Health Records (EHR). USCNet employs a Transformer-based multimodal fusion framework with CT-EHR attention and segmentation-guided attention modules for accurate classification. Moreover, a dynamic loss function is introduced to effectively balance the dual objectives of segmentation and classification. Experiments on an in-house kidney stone dataset show that USCNet demonstrates outstanding performance across all evaluation metrics, with its classification efficacy significantly surpassing existing mainstream methods. This study presents a promising solution for the precise preoperative classification of kidney stones, offering substantial clinical benefits. The source code has been made publicly available: https://github.com/fancccc/KidneyStoneSC.**

***Index Terms*— Urolithiasis Analysis, Clinical Information, Multimodal Learning, Dynamic Loss**

This work was supported by China Ministry of Education Humanities and Social Sciences Research-Planning Fund Project (No. 23YJA910007), Zhejiang Provincial Philosophy and Social Sciences Planning Project-Major Project (No. 23SYS12ZD), First Class Discipline of Zhejiang - A (Zhejiang University of Finance and Economics-Statistics), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617, 2022A1515110570), Innovation Teams of Youth Innovation in Science and Technology of High Education Institutions of Shandong Province (No. 2021KJ088). (Corresponding author: Yongquan Zhang and Hai Wang.).

C. Wang and X. Wan are with Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: wangcm@sribd.cn; wanxiang@sribd.cn).

Y. Zhang, S. Zhang, and Y. Wang are with Zhejiang University of Finance and Economics, Hangzhou 310018, China (e-mail: zyq@zufe.edu.cn; qq1442449344@gmail.com; 220110770228@zufe.edu.cn).

L. Liu is with Anhui University of Finance and Economics, Anhui 233000, China (e-mail: 20231441@aufe.edu.cn).

X. Li, J. Pan, and Y. Jiang are with The Second Affiliated Hospital of Chinese University of Hong Kong (Longgang District People's Hospital of Shenzhen), Shenzhen 518172, China (e-mail: xingzhi02459@gmail.com; jiexinp96@gmail.com; jackyamarelle@gmail.com).

H. Wang is with the School of Engineering and Energy and Harry Butler Institute, Murdoch University, Perth, WA6150, Australia (e-mail: Hai.Wang@murdoch.edu.au).

N. Li is with the School of Computer Science and Engineering, Macau University of Science and Technology, Macao 999078, China (e-mail: nnli@must.edu.mo).

A. Elazab is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: ahmed.elazab@yahoo.com).

I. Introduction

UROLITHIASIS, a widespread urinary tract disease, impacts millions globally, with its incidence notably increasing in industrialized nations [1], [2]. This rise is linked to modern lifestyle habits, including poor diet, inadequate hydration, sedentary lifestyle, and diets high in sodium or oxalate. The condition predominantly affects individuals aged 20–40, with a higher occurrence in males, possibly due to the influence of testosterone on stone formation [3]. Urolithiasis can be divided into infectious and non-infectious types [4]. Infectious stones, mainly composed of magnesium ammonium phosphate, form in alkaline urine and are often associated with recurrent urinary tract infections caused by urease-producing bacteria [5]. Conversely, non-infectious stones, such as those formed from calcium oxalate, phosphate, uric acid, and cystine, are primarily due to metabolic disorders, dehydration, or genetic predispositions [6], though they can also lead to secondary infections in cases of urinary obstruction or retention. Accurate preoperative prediction of stone composition is vital for selecting appropriate treatment strategies and advising lifestyle changes to prevent recurrence [7]. Traditional diagnostic methods typically depend on postoperative chemical stone analysis [8] or diagnoses inferred from clinical symptoms and laboratory results [9]. These methods fall short of providing real-time assessments of stone composition heterogeneity, spatial distribution, and dynamic changes, while overlooking important factors such as stone size, location, and interactions with surrounding tissues. Although computed tomography (CT) imaging is the gold standard for diagnosing urolithiasis [10]–[13], offering precise stone localization and aiding treatment decisions [14], it still faces challenges in accurately classifying stone types. Consequently, achieving preoperative classification of in vivo urinary stones using CT imaging remains a significant challenge.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

Over the last decade, the rapid progress in deep learning has led to remarkable advancements in the early detection of urinary stones [15]. Convolutional neural networks (CNNs), particularly those from the ResNet series [16], have significantly enhanced the accuracy of identifying stone features in CT images due to their advanced feature extraction capabilities. ResNet addresses the vanishing gradient issue often encountered in deep network training by incorporating residual connections, which allow for the creation of deeper networks, thereby improving model precision and performance. Several studies have explored the use of CNNs for classifying urinary stones through CT imaging. For instance, Hines *et al.* [17] and Shen *et al.* [18] examined CNN applications in the analysis of microscopic CT images to increase the automation and accuracy of stone classification. In related research, Asif *et al.* [19] introduced StoneNet, which utilizes the MobileNet architecture enhanced with global average pooling, batch normalization, dropout, and dense layers to reduce the number of parameters and increase model robustness. Similarly, Elton *et al.* [20] developed an automated system for detecting kidney stones and performing volumetric segmentation in non-contrast CT scans, achieving high sensitivity and specificity. Furthermore, Patro *et al.* [21] proposed a computer-aided diagnostic system that integrates convolution techniques based on the Kronecker product, which effectively minimizes redundancy in feature maps without overlapping convolutions. Cheng *et al.* [22] introduced ResGANet, incorporating modular group attention blocks to enhance feature representation in medical image classification and segmentation tasks while also reducing parameter complexity. Despite these advancements, relying solely on CT imaging poses significant challenges, particularly in accurately classifying stone types. Many stones, such as calcium oxalate and uric acid stones, have similar radiodensity on CT scans, making differentiation difficult.

In multimodal learning, Baharoon *et al.* proposed the HyMNet model, which enhances feature extraction and disease prediction accuracy by combining visual data from fundus images with clinical information on cardiovascular metabolic risk factors [23]. However, its reliance on simple feature concatenation limits its ability to fully explore deep correlations and interactions between modalities, potentially leading to redundant or lost information and affecting performance on complex datasets. To improve multimodal fusion, a Transformer-based mechanism can be introduced. Recent studies have demonstrated the effectiveness of this approach in medical applications: Yu *et al.* [24] developed ICH-PRNet, a novel cross-modal network for ICH prognosis prediction that effectively combines CT scans and clinical notes through an advanced joint-attention mechanism, while Yu *et al.* [25] created ICH-SCNet, an innovative framework leveraging CLIP-guided SAM technology for simultaneous segmentation and classification. The Transformer architecture [26] revolutionized deep learning by replacing traditional CNNs with self-attention, capturing global dependencies among sequence elements. Initially designed for textual data, the Transformer's applicability to other data types was limited due to its focus on discrete word sequences. To extend its use, Google Research developed the Vision Transformer (ViT) [27], which divides images into fixed-size patches and processes them as sequences through a Transformer encoder, opening new possibilities for multimodal data processing. Additionally, Hatamizadeh *et al.* [28]proposed the UNETR model, which combines the Transformers' global context modeling with U-Net's [29] precise localization, achieving superior performance in medical image segmentation. Recently, Xing *et al.* [30] introduced a hybrid masked image modeling method that significantly enhances 3D medical image segmentation accuracy.

Current methods for integrating segmentation with classification tasks fall into two main categories: segmentation-assisted classification and segmentation-classification multi-task frameworks. In the segmentation-assisted classification approach, Zhu *et al.* [31] introduced the SegPrompt model, which utilizes pretrained models that employ segmentation maps as prompts for classifying kidney stones. However, this method has several drawbacks: it does not incorporate segmentation during training, the segmentation maps are not directly linked to classification, and it depends heavily on pre-training. On the other hand, in the segmentation-classification multi-task framework, Saeed *et al.* [32] developed the TMSS network, which conducts cancer segmentation and survival prediction simultaneously. Although TMSS uses Transformers for comprehensive processing of multimodal data, it treats segmentation and classification as independent tasks without sharing features interactively. Furthermore, its fixed weighting mechanism for loss calculation restricts dynamic adjustments, leading to task imbalances that adversely affect overall performance.

Building on prior research, the accurate preoperative classification of urinary stones presents three significant challenges. First, single CT imaging falls short in capturing the patient-specific clinical context that characterizes infectious stones. Second, current segmentation-assisted methods only offer static spatial prompts, lacking the dynamic interaction necessary between classification and segmentation. Third, multi-task frameworks encounter inherent conflicts between the optimization objectives for segmentation and classification. To address these challenges, we propose a novel model, USCNet, designed for in vivo urinary stone classification. USCNet integrates Transformer-based multimodal fusion with segmentation-assisted classification, leveraging multimodal data from CT images and Electronic Health Records (EHR) [33]. As a multi-task multimodal framework, USCNet simultaneously performs segmentation and classification, enhancing the synergy between these processes. By incorporating clinical data and segmentation feature maps into the classification workflow via self-attention mechanisms, the model ensures improved accuracy and contextual understanding. The key contributions of USCNet are as follows:

1) To overcome the limitations of traditional unimodal methods, we developed the CT-EHR attention module, which enhances the fusion of imaging features with clinical context. This module dynamically integrates CT images with patient-specific clinical data at the feature level through cross-modal interaction mechanisms. By incorporating clinical background information, the model can more effectively distinguish infectious stones.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

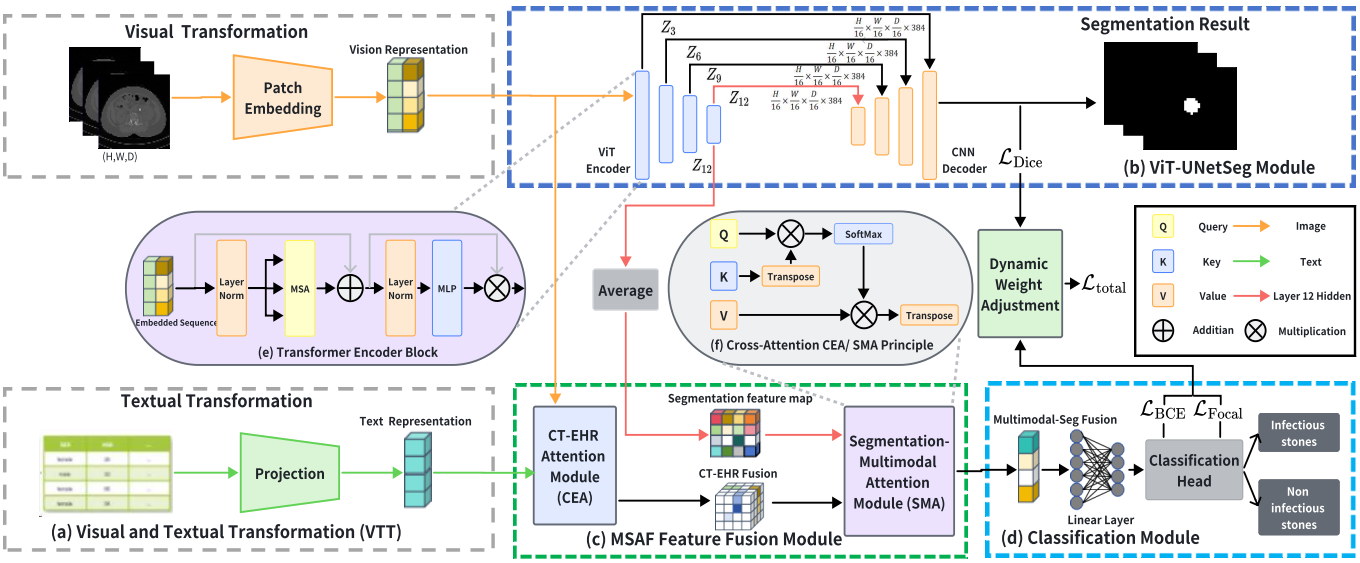

Fig. 1. The architecture of the USCNet: (a) the visual and textual transformation module, (b) the ViT-UNetSeg module, (c) the MSAF feature fusion module, (d) the classification module, (e) the Transformer encoder block, and (f) the cross-attention principle.

2) Unlike SegPrompt [31], which utilizes segmentation maps as static prompts, we introduce the Segmentation Multimodal Attention (SMA) module. The primary aim of this module is to enable the rich spatial information generated during segmentation to actively and dynamically inform classification decisions. By employing trainable attention mechanisms, the SMA module autonomously establishes semantic relationships between segmentation features and classification tasks, ensuring the model concentrates on the most discriminative anatomical structures for differentiating stone types.

3) In our segmentation-assisted classification framework, dependable early-stage segmentation features are crucial for effective classification outcomes. Unlike existing multi-task frameworks [23], [32] that rely on fixed-weight loss functions, we propose a dynamic loss function. This function adaptively adjusts the loss weights by continuously monitoring the performance metrics of each task in real-time. This approach allows for the coordinated optimization of both segmentation and classification tasks during training, offering a novel optimization paradigm for multi-task learning in complex scenarios.

The remainder of this paper is organized as follows: Section II details the proposed USCNet architecture, including the Visual and Textual Transformation Module, ViT-UNetSeg Module, and Multimodal-Segmentation Attention Fusion Module, with emphasis on the dynamic weight adjustment for total loss. Section III presents the comprehensive experimental study, covering: (1) data preparation and implementation details, (2) comparative experiments with state-of-the-art methods, and (3) ablation studies investigating key modules, hierarchical feature combinations, loss components, and adaptive loss functions. Finally, Section IV concludes the study.

II. Methods

The network architecture is effectively designed, as shown in Fig. 1, and includes four main components: the Visual and Textual Transformation (VTT) module (Section II-A), the segmentation module, the Multimodal-Segmentation Attention Fusion (MSAF) module (Section II-C), and the classification module. The VTT module preprocesses and transforms CT images and EHR data into a unified feature space by embedding CT images and projecting EHR data for a harmonized representation. The segmentation module (Section II-B) utilizes a UNETR [28] structure with a ViT [27] encoder to segment the visual representation output from the VTT module, thereby obtaining segmentation results. Our primary objective is to assess the infectivity of stones, and the Transformer module within the UNETR [28] architecture excels at pinpointing the spatial locations of stones in images. The MSAF module comprises the CEA and SMA modules, which integrate and refine features by selectively emphasizing clinically relevant information derived from images and EHR data. The classification module primarily includes the MSAF module, a linear layer, and a classification head. Its main function is to derive classification results from the MSAF module's output via the linear layer and classification head. Furthermore, the overall loss function for segmentation and classification is dynamically adjusted by the loss weighting module (Section II-D) to ensure balanced optimization of both tasks.

A. Visual and Textual Transformation Module

In this module, CT images undergo preprocessing through the Visual Transformation process. Specifically, the 3D image with dimensions $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{H}\times\boldsymbol{W}\times\boldsymbol{D}}$ is reshaped into a sequence of flat 2D patches $\boldsymbol{x}_p \in \mathbb{R}^{\boldsymbol{N}\times(\boldsymbol{P}^3)}$. Here, $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{D}$ denote the width, height, and depth of the CT images, respectively, while $\boldsymbol{P}$ signifies the dimension of each patch. The total number of patches $\boldsymbol{N}$ is calculated as follows:

$$\boldsymbol{N} = \frac{\boldsymbol{W} \times \boldsymbol{H} \times \boldsymbol{D}}{\boldsymbol{P}^3}, \tag{1}$$

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

where $\boldsymbol{P} \times \boldsymbol{P} \times \boldsymbol{P}$ represents the dimensions of each patch. To preserve the spatial relationships among segments, position embedding [27] is added to each patch after partitioning.

Simultaneously, the EHR data is transformed into a comparable feature space through the Textual Transformation. This process results in a linear feature representation with dimensions $[\boldsymbol{B}, \boldsymbol{E}]$, where $\boldsymbol{B}$ and $\boldsymbol{E}$ represent the batch and embedding sizes, respectively.The primary aim of this module is to harmonize heterogeneous data modalities, specifically high-dimensional CT images and structured EHR text, into a unified format. This harmonization is crucial for facilitating seamless integration and analysis in the subsequent phases of multimodal analysis. This preparatory step lays the foundation for a comprehensive cross-modal examination later in the multimodal analytical workflow.

Remark 1. The Visual and Textual Transformation Module bridges the modality gap between medical imaging and clinical data through innovative 3D patch embedding and synchronized positional encoding, preserving critical anatomical relationships while enabling direct cross-modal feature alignment.

B. The ViT-UNetSeg Module

The segmentation module is tailored for 3D medical image segmentation, integrating a ViT-based encoder with a U-Net style decoder to extract spatial features at multiple scales and produce accurate segmentation masks. The encoder initially divides the 3D input volume into non-overlapping patches of size $\boldsymbol{P} \times \boldsymbol{P} \times \boldsymbol{P}$, where $\boldsymbol{P} = 16$. These patches are then flattened and processed through 12 Transformer layers, each possessing a hidden size of 384, to capture hierarchical global contextual information. Intermediate features from the encoder's layers $\mathbf{Z}_3$, $\mathbf{Z}_6$, $\mathbf{Z}_9$, and $\mathbf{Z}_{12}$ (the final layer) are extracted to represent multi-scale spatial information.

The decoder employs a hierarchical upsampling structure, consisting of multiple UnetrUpBlock and UnetrPrUpBlock layers [28]. Each upsampling block incrementally increases the spatial resolution of the feature maps while reducing the number of channels. Skip connections are employed to merge multi-scale features from the encoder ($\mathbf{Z}_3$, $\mathbf{Z}_6$, $\mathbf{Z}_9$, and $\mathbf{Z}_{12}$) with the corresponding decoder layers. This approach ensures the effective utilization of both local details and global context for precise 3D segmentation. Finally, the UnetOutBlock [28] generates the 3D segmentation mask by transforming the high-resolution feature maps into the required number of output channels. This modular design enables the model to harness the global feature extraction strengths of the Transformer along with the local feature refinement capabilities of the U-Net.

Remark 2. This segmentation module innovatively combines a ViT encoder with a U-Net decoder through 3D patch embedding and hierarchical Transformer features. The progressive upsampling with skip connections achieves synergistic optimization of global context modeling and local detail preservation, inheriting both the Transformer's global perception and the U-Net's local refinement capabilities.

C. Multimodal-Segmentation Attention Fusion Module

The MSAF module comprises CEA and SMA modules, arranged sequentially and operating on a cross-attention mechanism. Initially, the CEA module combines the visual and textual representations from the VTT module to create the CT-EHR Fusion Representation (CEFR). Subsequently, the SMA module merges the segmentation feature map, sourced from the 12th transformer block of the ViT encoder, with the CEFR to generate the Multimodal-Segmentation Fusion Representation (MSFR). This MSFR output is then used for classification.

CT and EHR Attention: Within the MSAF module, the CEA submodule plays a crucial role in integrating CT image features with EHR data. It uses the vision representation as queries ($\boldsymbol{Q}_1$) and the EHR data as keys ($\boldsymbol{K}_1$) and values ($\boldsymbol{V}_1$). Based on the attention mechanism described in [26], the CEFR is computed as follows:

$$\text{Attention}(\boldsymbol{Q}_1, \boldsymbol{K}_1, \boldsymbol{V}_1) = \text{softmax}\left(\frac{\boldsymbol{Q}_1 \boldsymbol{K}_1^T}{\sqrt{\boldsymbol{d}_{k1}}}\right) \boldsymbol{V}_1, \quad (2)$$

where $\boldsymbol{d}_{k_1}$ is the dimension of $\boldsymbol{K}_1$. This setup allows the model to focus on image patches informed by EHR data, ensuring that the analysis is enriched with patient-specific context.

Segmentation Multimodal Attention: The SMA submodule enhances cross-modal data integration by combining the segmentation feature map with CEFR. It uses the segmentation feature map as queries ($\boldsymbol{Q}_2$) and the CEFR as keys ($\boldsymbol{K}_2$) and values ($\boldsymbol{V}_2$). The MSFR is calculated as follows:

$$\text{Attention}(\boldsymbol{Q}_2, \boldsymbol{K}_2, \boldsymbol{V}_2) = \text{Softmax}\left(\frac{\boldsymbol{Q}_2 \boldsymbol{K}_2^T}{\sqrt{\boldsymbol{d}_{k_2}}}\right) \boldsymbol{V}_2, \quad (3)$$

where $\boldsymbol{d}_{k_2}$ is the dimension of $\boldsymbol{K}_2$. This fusion process enhances the model's ability to focus on significant areas of the image identified by segmentation features. Consequently, the SMA ensures that tasks such as classification or subsequent analyses prioritize the most relevant regions of the image, informed by both segmentation and the broader multimodal context.

Remark 3. The MSAF module achieves triple fusion of imaging, clinical data, and anatomical segmentation through cascaded CEA and SMA attention mechanisms. The CEA submodule establishes patient-specific context by treating CT features as queries and EHR as keys and values, while the SMA submodule further guides multimodal representation to focus on critical regions using segmentation features as queries, forming a hierarchical cross-modal attention architecture.

D. Dynamic Weight Adjustment for Total Loss

The USCNet architecture employs a composite loss function specifically designed for its dual tasks of segmentation and classification. This integrated loss function consists of three components: Dice loss [34] for segmentation, and binary cross-entropy (BCE) loss coupled with focal loss [35] for

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

classification. The overall loss is calculated as a dynamically weighted sum of these components, facilitated by an innovative automatic optimization strategy. This approach ensures a balanced focus between segmentation and classification tasks, with adjustments based on their respective performance metrics.

Dice Loss is used for the segmentation task, and its formula is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}, \tag{4}$$

where p_i represents the predicted value and g_i is the ground truth.

BCE Loss addresses the classification task, computed as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{\boldsymbol{N}} \sum_i \left[g_i \log(p_i) + (1 - g_i) \log(1 - p_i) \right], \tag{5}$$

where p_i is the predicted probability, g_i is the ground truth label, and N is the number of samples.

Focal Loss is employed to tackle class imbalance in classification, defined as:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{\boldsymbol{N}} \sum_i \left[g_i (1 - p_i)^\gamma \log(p_i) + (1 - g_i) p_i^\gamma \log(1 - p_i) \right], \tag{6}$$

where γ is the focusing parameter that adjusts the weight of hard and easy samples.

The total loss function is expressed as:

$$\mathcal{L}_{\text{total}} = \omega_{\text{Dice}} \mathcal{L}_{\text{Dice}} + \omega_{\text{Class}} (\lambda \mathcal{L}_{\text{BCE}} + (1 - \lambda) \mathcal{L}_{\text{Focal}}), \tag{7}$$

where ω_{Dice} and ω_{Class} denote the weights assigned to the Dice loss and the composite classification losses, respectively. λ is a predetermined constant balancing the BCE and focal losses, such as a 1:10 ratio.

Weight Adjustment Strategy: As shown in Algorithm 1, when the Dice score S_{Dice} is 0.8 or lower, the weights are dynamically adjusted: $\omega_{\text{Dice}} = 1 - S_{\text{Dice}}$ and $\omega_{\text{Class}} = S_{\text{Dice}}$. The classification loss weights are divided into $\lambda \mathcal{L}_{\text{BCE}}$ and $(1 - \lambda)\mathcal{L}_{\text{Focal}}$. When the Dice score exceeds 0.8, fixed weights prioritize classification: $\omega_{\text{Dice}} = 0.2$ and $\omega_{\text{Class}} = 0.8$, maintaining the division of classification loss components.

Algorithm 1 Dynamic weight adjustment for USCNet.

Initialize $\omega_{\text{Dice}} = 0.8$, $\lambda = 0.1$, $\{\omega_{\text{Class}} = \omega_{\text{BCE}} + \omega_{\text{Focal}}\}$
for each training epoch **do**
  Compute Dice score S_{Dice} on validation set
  if $S_{\text{Dice}} \leq 0.8$ **then**
    $\omega_{\text{Dice}} \leftarrow 1 - S_{\text{Dice}}$
    $\omega_{\text{BCE}} \leftarrow \lambda S_{\text{Dice}}$
    $\omega_{\text{Focal}} \leftarrow S_{\text{Dice}} - \omega_{\text{BCE}}$ $\{(1 - \lambda) S_{\text{Dice}}\}$
  else
    $\omega_{\text{Dice}} \leftarrow 0.2$
    $\omega_{\text{BCE}} \leftarrow \lambda \times 0.8$ $\{0.08\}$
    $\omega_{\text{Focal}} \leftarrow 0.8 - \omega_{\text{BCE}}$ $\{0.72\}$
  end if
  $\mathcal{L}_{\text{total}} = \omega_{\text{Dice}} \mathcal{L}_{\text{Dice}} + \omega_{\text{BCE}} \mathcal{L}_{\text{BCE}} + \omega_{\text{Focal}} \mathcal{L}_{\text{Focal}}$
end for

This dynamic weight adjustment strategy enables the model to shift focus towards enhancing classification performance as segmentation accuracy improves. Initially, a higher segmentation loss weight encourages rapid segmentation convergence. As segmentation accuracy improves, the focus shifts to classification. By recalibrating these dynamic weights during each training epoch, the model maintains a balanced optimization of both tasks, significantly enhancing overall efficacy.

Remark 4. This module proposes a dynamic task-balancing strategy that intelligently adjusts segmentation and classification loss weights based on Dice scores. It prioritizes segmentation convergence initially, then automatically shifts focus to classification when reaching the threshold, establishing a "segmentation-first-classification-later" progressive training paradigm. This design resolves optimization conflicts in multi-task learning inherent to fixed-weight approaches.

III. Experiments

A. Data Preparation and Preprocessing

We collected data from 642 patients who underwent CT imaging followed by surgical treatment for urolithiasis. This dataset includes preoperative urological CT images in DICOM format, preoperative complete blood count results, urinalysis, and postoperative stone composition analysis. Overall, the dataset comprises 1,355 CT images of kidney stones. The postoperative stone composition analysis, regarded as the gold standard, was used to classify the stones into infectious and non-infectious categories. Stone masks were annotated by four experienced urologists using 3D Slicer software (version 5.2.2) based on surgical outcomes, with the accuracy of these annotations validated by two experienced radiologists.

This study utilized seven key clinical variables extracted from preoperative EHR: **1) Age, 2) Gender, 3) Blood Leukocyte Count, 4) Serum Creatinine, 5) Urine Leukocyte Count, 6) Urine pH,** and **7) Stone Location**. These variables were chosen based on their clinical significance and availability in routine preoperative evaluations. Notably, all variables, including stone location derived from CT imaging, were collected exclusively during the preoperative period. The ground truth labels, infectious versus non-infectious stones, were determined through postoperative composition analysis. This distinction between preoperative features and postoperative labels was integral to the study's design, ensuring a clear temporal separation that effectively mitigates the risk of data leakage. To further enhance the dataset's integrity, variables with substantial missing data or those not routinely assessed during preoperative care, such as urine culture results, were deliberately excluded to minimize selection bias and maintain consistency in clinical applicability.

The CT images were initially resampled to a uniform resolution of 1 mm³ to ensure consistency across all data. Given that a single patient could have multiple stones, images were cropped to isolate each stone region. First, a region of interest (ROI) was identified to locate each stone. The stones were then categorized as infectious or non-infectious, and these regions were extracted into cubes measuring 48 mm on each side. Only cubes containing stones with a volume of at

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

TABLE I
COMPARISONS WITH SOTA METHODS (BEST IN BOLD, SECOND-BEST UNDERLINED).

No.	Methods	Classification					Segmentation		
		Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	Dice↑(%)	IoU↑(%)	HD95↓(mm)
1	ResNet34 [16]	56.25±2.14	10.43±3.51	5.50±2.82	44.90±3.21	52.75±2.30	-	-	-
2	ResNet50 [16]	62.13±1.82	28.96±2.92	19.26±2.55	58.33±2.79	55.03±2.12	-	-	-
3	ResNet101 [16]	63.97±1.72	38.75±2.64	28.44±2.34	60.47±2.42	58.05±1.91	-	-	-
4	StoneNet [19]	59.93±2.01	57.22±2.25	49.69±2.64	40.07±3.52	50.04±2.46	-	-	-
5	SegPrompt [31]	61.27±1.96	46.94±2.45	42.20±2.40	52.87±2.81	58.52±2.01	-	-	-
6	nnUnet [36]	-	-	-	-	-	**86.35±0.41**	**75.68±0.41**	**3.50±0.35**
7	UNETR [28]	-	-	-	-	-	82.27±1.02	69.18±1.25	3.88±0.51
8	ResGANet [22]	62.50±2.02	47.17±2.82	44.31±2.62	46.96±3.11	56.37±2.36	83.85±1.22	72.08±1.22	4.80±0.45
9	TMSS [32]	74.63±1.84	57.67±2.52	43.12±2.85	87.04±2.22	81.54±1.73	79.13±1.43	65.43±1.44	6.20±0.60
10	HyMNet [23]	85.66±1.55	82.81±1.95	86.23±1.46	79.66±2.53	85.76±1.35	77.14±1.52	62.67±1.55	5.50±0.75
⋆	**USCNet (Ours)**	**91.91±1.42**	**90.00±1.74**	**90.83±0.94**	**89.19±3.13**	**91.73±1.30**	83.98±1.09	72.12±1.42	3.73±0.82

least 0.06 mm³ were selected for further analysis, resulting in a total of 2,113 stone image cases.

To ensure robust evaluation, the dataset was divided using a 5-fold cross-validation approach. Before being input into the network, the CT images underwent window leveling adjustment to the range of (-400, 2000) Hounsfield units, followed by normalization. Meanwhile, continuous variables in the clinical dataset were standardized, and categorical variables were converted into a one-hot encoded format.

B. Implementation Details

Our experiments were conducted on a 48G NVIDIA A40 GPU running CUDA Version 12.2, using Python 3.11 and PyTorch 2.4. Using an input size of 48^3 voxels and a batch size of 128, the training process consumed 11.52 GB of GPU memory at its peak. Each training fold required 0.92 hours to complete, equivalent to 0.46 GPU-hours. To ensure robust model evaluation, we employed a five-fold cross-validation strategy. This involved dividing the dataset into five subsets, with each subset serving once as the validation set while the remaining subsets were used for training. We maintained consistent class distributions across all folds through stratified splitting.

The model was trained for 200 epochs with a batch size of 128 and an initial learning rate of 0.001. We utilized the Adam optimizer, configured with a weight decay of 0.001 and betas set to (0.9, 0.99). These parameters controlled the exponential decay rates for the first and second moment estimates, aiding in momentum handling and adaptive learning rate adjustments. To enhance training stability, we implemented the ReduceLROnPlateau learning rate scheduler, which reduces the learning rate when the validation loss plateaus. The scheduler was set with a reduction factor of 0.1 and a patience of 10 epochs, ensuring adjustments only when necessary.

The segmentation module employed a ViT architecture based on the UNETR framework [28], consisting of 12 layers with 12 attention heads each. The adaptive loss function began with initial weights of 0.2 for the classification loss and 0.8 for the Dice loss. We set $\lambda = 0.1$, resulting in weights of 0.02 for the BCE loss and 0.18 for the focal loss, which were adjusted based on segmentation performance. To address class imbalance, the focal loss was modulated by parameters γ and α, set at 2 and 0.25, respectively, to focus on hard-to-classify samples and balance positive and negative sample weights. This configuration was designed to enhance model performance on imbalanced datasets.

C. Comparative Experiments

1) Comparative Methods: We evaluated the proposed USCNet model against several traditional ResNet architectures [16], including ResNet34 [16], ResNet50 [16], and ResNet101 [16], to assess its classification performance. Additionally, we compared USCNet with StoneNet [19] and SegPrompt [31], two prominent models in the field of stone classification. These benchmark models were used to demonstrate USCNet's effectiveness in classification tasks.

Furthermore, we assessed USCNet's segmentation capabilities by comparing it with nnU-net [36], a well-regarded automated medical image segmentation framework based on U-Net, known for its strong generalization and efficiency in medical image segmentation. We also included UNETR [28], an advanced Transformer-based benchmark model for medical image segmentation, to provide comparative performance analysis. Furthermore, we conducted comprehensive comparisons with ResGANet [22], TMSS [32], and HyMNet [23] in terms of both classification and segmentation performance. Notably, ResGANet employs independent modules for classification and segmentation without exploiting potential inter-task relationships, while both TMSS and HyMNet are designed as multi-task, multi-modal fusion models that can simultaneously handle classification and segmentation tasks through multi-modal data integration. It should be noted that since HyMNet was originally designed for survival prediction tasks, we adapted its output layer structure to accommodate classification tasks. Through these comparative experiments, we have comprehensively demonstrated USCNet's superior

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

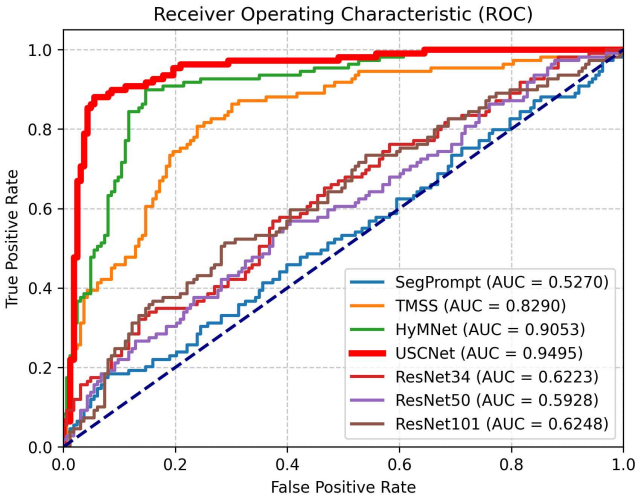

Fig. 2. ROC curve comparison of diagnostic accuracy between USCNet and established methods for infectious urolithiasis classification.

performance in jointly handling both classification and segmentation tasks.

2) Evaluation Metrics: We employed five key evaluation metrics to assess the classification performance of our proposed model in comparison to state-of-the-art approaches: Accuracy (Acc), F1 Score (F1), Recall (Rec), Precision (Pre), and Area Under the Curve (AUC). While Acc measures the overall proportion of correct predictions, it can be misleading in imbalanced datasets. In such cases, Pre and Rec are particularly valuable, as Pre evaluates the Acc of positive predictions, and Rec assesses the model's ability to identify all actual positives. The F1, which balances Pre and Rec, is especially useful for optimizing performance in imbalanced settings.

To evaluate segmentation accuracy, we employed three key metrics: the Dice Coefficient (Dice), Intersection over Union (IoU), and the 95th percentile Hausdorff Distance (HD95). The Dice Coefficient and IoU assess the overlap between predicted and actual segmentation masks, with higher values reflecting improved performance. The HD95 measures the boundary distance between the predicted and ground truth contours, where lower values signify greater boundary delineation accuracy. Collectively, these metrics provide a robust and comprehensive framework for assessing model performance across both segmentation and classification tasks.

3) Comparative Results: The results, presented in Table I, demonstrate that USCNet surpasses all comparative models in classification tasks across various metrics. Among the ResNet variants, the highest accuracy observed was 63.97%, with ResNet101 achieving the best performance in this family. StoneNet achieved an Acc of 59.93%, similar to the ResNet models, whereas SegPrompt reported an Acc of 61.27%. ResGANet, on the other hand, achieved 62.50% Acc in classification tasks. These relatively low Acc rates stem from the unique challenges associated with in vivo stone classification. The subtle variations in stone shape and size, coupled with the absence of color information in CT scans, hinder feature extraction and classification Acc. Consequently, single-modal models like StoneNet and SegPrompt, designed for in vitro stone classification, struggle with in vivo tasks due to their reliance solely on CT images.

In contrast, when comparing USCNet with state-of-the-art multi-modal methods, significant advancements are clear. While methods such as HyMNet and TMSS improved Acc to 85.66% and 74.63%, respectively, they typically use only CT images and EHR data, integrating them through simple concatenation. USCNet, however, incorporates segmentation feature maps with multi-modal inputs using a dual self-attention module for feature fusion, markedly enhancing classification Acc. By developing a dynamic loss function, USCNet adaptively balances segmentation and classification losses based on task-specific metrics, ensuring that segmentation features effectively support classification tasks and bolstering the model's feature learning capabilities. Ultimately, USCNet achieved an Acc of 91.91%, representing improvements of 27.94%, 31.98%, 30.64%, 29.41%, 17.28%, and 6.25% over ResNet architectures, StoneNet, SegPrompt, ResGANet, TMSS, and HyMNet, respectively. The performance of USCNet was further validated through DeLong's test, which confirmed a statistically significant improvement in AUC compared to the second-best method, HyMNet ($p < 0.001$). Figure 2 illustrates the ROC curve, where USCNet's proximity to the top-left corner highlights its excellent performance in distinguishing between positive and negative samples. The high AUC further validates the model's Acc and reliability. Figure 3 displays the confusion matrix for USCNet, showing a relatively low number of misclassifications, thus underscoring the model's robust performance.

Beyond these quantitative metrics, USCNet provides significant clinical value by predicting stone infectivity through the integration of preoperative imaging and clinical data. This capability offers crucial decision support for perioperative management and may help reduce postoperative complication rates. Importantly, USCNet's application extends beyond high-resource clinical settings - its implementation in underserved areas could assist less experienced clinicians in making accurate diagnoses, thereby improving healthcare accessibility. This democratization of diagnostic capability is particularly valuable in regions with significant disparities in medical resources.

In the segmentation task, USCNet achieved a Dice score of 85.64%, an IoU of 72.12%, and an HD95 of 3.73 mm, delivering strong overall performance, second only to nnU-net. This outcome reflects the core principle of our adaptive weighting strategy: the model prioritizes extracting effective segmentation features to support classification rather than maximizing segmentation accuracy alone. Once the segmentation quality reaches a stable level sufficient to aid classification, the optimization focus shifts dynamically toward enhancing classification performance. Notably, USCNet surpassed all other comparative methods except nnU-net in HD95 boundary accuracy, indicating its segmentation masks provide more precise boundary information. By maintaining robust segmentation quality, USCNet significantly improved classification performance, effectively demonstrating the strength of our segmentation-assisted classification approach.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

TABLE II
ABLATION STUDIES OF THE PROPOSED DYNAMIC LOSS FUNCTION WITH DIFFERENT COMPONENTS.

Loss Function Components					Classification					Segmentation		
CT	Clinical	CEA	SMA	DW	Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	Dice↑(%)	IoU↑(%)	HD95↓(mm)
-	✓	-	-	-	63.97±3.38	59.17±3.26	65.14±4.37	54.20±5.25	64.16±3.82	-	-	-
✓	-	-	-	-	-	-	-	-	-	82.27±1.02	69.18±1.25	3.88±0.51
✓	✓	-	-	-	70.56±4.88	66.90±5.62	74.33±5.27	60.90±6.15	70.16±4.84	78.33±1.73	64.30±1.73	7.42±0.87
✓	✓	✓	-	-	89.71±1.86	87.16±2.27	87.16±1.69	87.16±3.39	89.28±1.79	82.19±0.93	69.54±0.93	4.31±0.30
✓	✓	✓	✓	-	89.34±1.38	86.64±1.52	86.24±2.06	87.04±1.78	88.82±1.40	81.44±0.95	68.34±0.95	4.03±0.43
✓	✓	✓	✓	✓	**91.91±1.42**	**90.00±1.74**	**90.83±0.94**	**89.19±3.13**	**91.73±1.30**	**83.98±1.09**	**72.12±1.42**	**3.73±0.82**

TABLE III
ABLATION STUDIES OF SEGMENTATION-ASSISTED CLASSIFICATION WITH DIFFERENT FEATURE COMBINATIONS.

Feature Comb.				Classification					Segmentation		
$\mathbf{Z}_3$	$\mathbf{Z}_6$	$\mathbf{Z}_9$	$\mathbf{Z}_{12}$	Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	Dice↑(%)	IoU↑(%)	HD95↓(mm)
-	-	-	✓	**91.91±1.42**	**90.00±1.74**	**90.83±0.94**	**89.19±3.13**	**91.73±1.30**	**83.98±1.09**	**72.12±1.42**	**3.73±0.82**
-	-	✓	-	90.44±1.54	88.29±1.91	89.91±1.51	86.73±3.22	90.35±1.44	81.93±0.99	69.49±0.99	4.66±0.56
-	✓	-	-	90.07±2.00	87.78±2.34	88.99±2.02	86.61±3.50	89.89±1.93	82.29±0.96	69.71±0.96	4.18±0.49
✓	-	-	-	87.87±2.10	84.93±2.43	85.32±2.14	84.55±3.02	87.45±2.05	79.98±1.07	66.67±1.07	5.33±0.25
✓	✓	-	-	89.71±1.82	87.39±2.13	88.99±1.14	85.84±3.75	89.59±1.64	82.05±1.02	69.47±1.02	4.59±0.43
✓	-	✓	-	89.71±2.01	87.27±2.38	88.07±1.43	86.49±3.77	89.44±1.87	82.24±0.92	69.58±0.92	4.40±0.19
✓	-	-	✓	90.07±1.72	87.78±2.02	88.99±1.44	86.61±3.47	89.89±1.59	82.32±0.89	69.83±0.89	4.10±0.27
-	✓	✓	-	89.34±1.72	87.00±2.10	88.99±1.55	85.09±3.75	89.28±1.58	82.45±1.02	69.91±1.02	4.40±0.26
-	✓	-	✓	89.71±2.00	87.16±2.45	87.16±2.62	87.16±3.44	89.28±2.02	81.61±1.05	68.73±1.05	4.77±0.50
-	-	✓	✓	90.44±1.67	88.39±1.94	90.83±1.47	86.09±3.21	90.50±1.53	82.60±0.81	70.22±0.81	4.20±0.46
✓	✓	✓	-	89.71±1.53	87.27±1.69	88.07±1.02	86.49±2.67	89.44±1.40	80.80±1.08	67.43±1.08	4.32±0.39
✓	✓	-	✓	90.44±1.50	88.29±1.76	89.91±1.35	86.73±3.21	90.35±1.37	82.34±0.98	69.85±0.98	4.12±0.41
✓	-	✓	✓	89.34±2.16	86.76±2.65	87.16±2.09	86.36±3.85	88.98±2.10	82.31±0.96	69.85±0.96	4.18±0.43
-	✓	✓	✓	89.71±1.77	87.27±2.07	88.07±1.48	86.49±3.58	89.44±1.64	81.28±0.99	68.12±0.99	4.61±0.44
✓	✓	✓	✓	89.71±1.99	87.16±2.45	87.16±2.23	87.16±3.51	89.28±1.97	81.90±1.10	69.17±1.10	4.30±0.65

TABLE IV
ABLATION STUDIES OF DIFFERENT LOSS FUNCTIONS FOR SEGMENTATION-ASSISTED CLASSIFICATION.

Loss Function			Classification					Segmentation		
$\mathcal{L}_{\text{BCE}}$	$\mathcal{L}_{\text{Focal}}$	$\mathcal{L}_{\text{Dice}}$	Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	Dice↑(%)	IoU↑(%)	HD95↓(mm)
✓	–	✓	91.54±1.55	89.50±1.85	89.91±1.72	89.09±3.01	91.27±1.49	82.43±1.25	70.11±1.58	4.19±0.49
–	✓	✓	91.18±1.20	88.68±1.35	86.24±2.38	91.26±1.75	90.36±1.30	83.02±2.17	70.96±2.32	4.02±1.59
✓	✓	✓	**91.91±1.42**	**90.00±1.74**	**90.83±0.94**	**89.19±3.13**	**91.73±1.30**	**83.98±1.09**	**72.12±1.42**	**3.73±0.82**

TABLE V
EFFECT OF THE ADAPTIVE LOSS FUNCTIONS ON THE CLASSIFICATION AND SEGMENTATION PERFORMANCES OF THE USCNET MODEL.

Class. / Seg.	Classification					Segmentation		
	Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	Dice↑(%)	IoU↑(%)	HD95↓(mm)
0.1 / 0.9	82.35±2.23	77.14±2.59	74.31±2.66	80.20±3.84	81.02±2.11	78.86±1.18	66.67±1.18	5.66±0.35
0.3 / 0.7	86.40±1.39	83.41±1.36	85.32±1.65	81.58±3.18	86.22±1.14	81.96±1.11	69.00±1.11	4.38±0.64
0.5 / 0.5	89.34±1.38	86.64±1.52	86.24±2.06	87.04±1.78	88.82±1.40	81.44±0.95	68.34±0.95	4.03±0.43
0.7 / 0.3	89.71±1.90	87.27±2.29	88.07±1.46	86.49±3.65	89.44±1.78	82.04±1.20	69.25±1.20	4.18±0.58
0.9 / 0.1	90.44±1.67	88.39±1.93	90.83±1.21	86.09±3.67	90.50±1.47	82.11±1.02	69.47±1.02	4.78±0.31
Auto (Ours)	**91.91±1.42**	**90.00±1.74**	**90.83±0.94**	**89.19±3.13**	**91.73±1.30**	**83.98±1.09**	**72.12±1.42**	**3.73±0.82**

D. Subgroup Analysis

To evaluate the clinical applicability of USCNet across diverse patient populations, a detailed subgroup analysis was performed based on demographic and clinical characteristics. As summarized in Table VI, the model consistently demonstrated strong performance across all subgroups, with several significant trends emerging.

USCNet achieved consistently high accuracy across age groups, with results of 91.80% for patients under 40 years, 90.11% for those aged 40–60 years, and 91.18% for patients over 60 years. Notably, the youngest cohort achieved the highest AUC of 96.99%, indicating the model's ability to effectively extract features despite age-related anatomical variations. Similarly, both male and female patients showed comparable performance, with no evidence of gender-based bias affecting classification accuracy.

A particularly compelling finding was the model's exceptional performance in detecting infectious stones, achieving

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

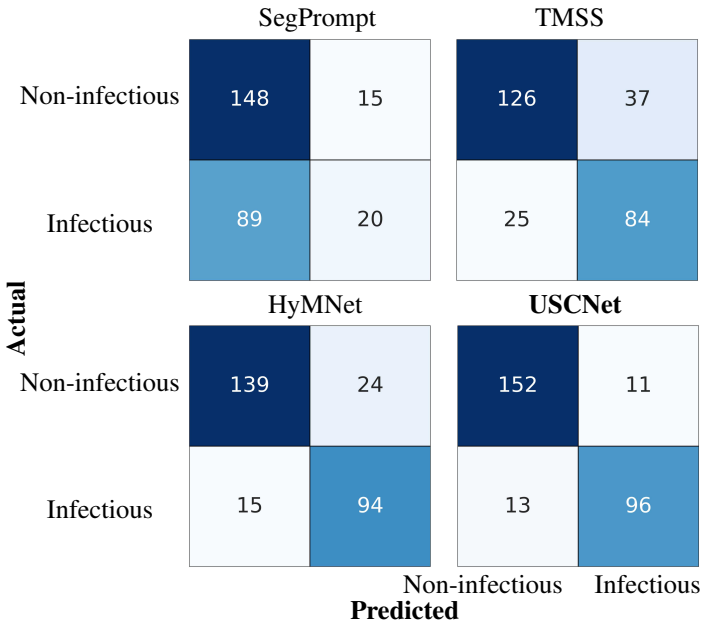

Fig. 3. Confusion matrices comparison of USCNet versus multimodal baseline methods, quantitatively evaluating classification performance for infectious urolithiasis diagnosis.

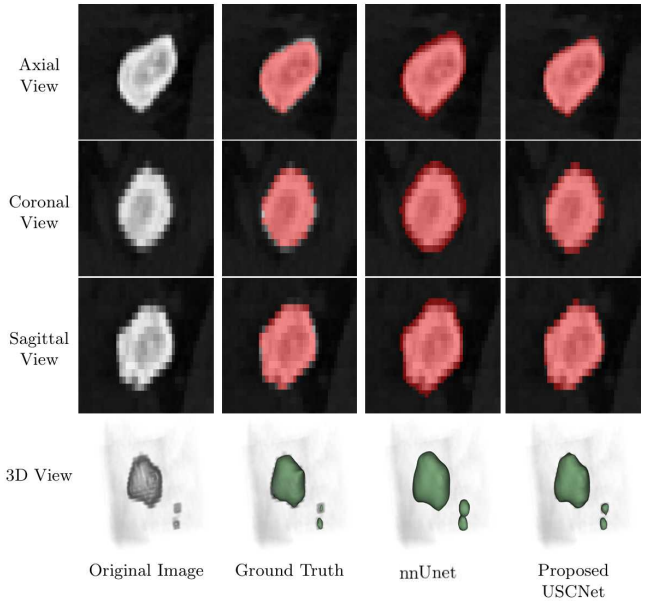

Fig. 4. Comparison of 3D multi-planar segmentation results (original images, ground truth annotations, nnUnet and USCNet outputs).

96.40% accuracy and an AUC of 99.63%, significantly outperforming its results for non-infectious stones. This highlights USCNet's ability to identify infection-related stone characteristics, a critical factor for guiding treatment decisions such as antibiotic selection.

In terms of anatomical stone location, the model performed exceptionally well for right ureter stones and achieved perfect AUC scores for stones located in the left kidney, right kidney, and other kidney-specific locations. These results may reflect anatomical differences and distinct imaging features of stones in various segments of the urinary tract, offering valuable insights for clinical application in diverse urological settings.

TABLE VI
PERFORMANCE ANALYSIS OF USCNET ACROSS DIFFERENT PATIENT SUBGROUPS.

Subgroup	Performance					N
	Acc↑(%)	F1↑(%)	Rec↑(%)	Pre↑(%)	AUC↑(%)	
Age Groups						
<40 years	91.80	87.80	94.74	81.82	96.99	61
40-60 years	90.11	88.00	86.84	89.19	94.09	91
>60 years	91.18	88.00	91.67	84.62	94.32	34
Gender						
Male	90.98	87.91	93.02	83.33	95.70	122
Female	90.62	88.00	84.62	91.67	94.33	64
Stone Type						
Non-infectious	87.58	84.38	83.08	85.71	91.52	161
Infectious	96.40	95.45	95.45	95.45	99.63	111
Stone Location						
Left Ureter	89.53	85.71	79.41	93.10	93.04	86
Right Ureter	98.11	97.56	100.00	95.24	96.97	53
Ureter	87.30	87.10	90.00	84.38	90.00	63
Left Kidney	93.33	80.00	66.67	100.00	100.00	15
Right Kidney	88.24	90.91	83.33	100.00	100.00	17
Kidney	93.75	80.00	100.00	66.67	100.00	16
Bladder	90.00	90.91	100.00	83.33	96.00	10
Urethra	85.71	66.67	100.00	50.00	100.00	7

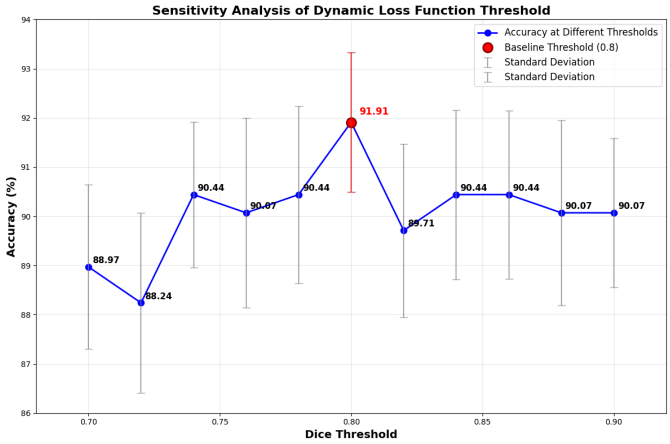

Fig. 5. Empirical validation of the dynamic loss threshold through systematic sensitivity analysis.

E. Ablation Study

1) Ablation Study on Key Modules: This ablation study examines the individual contributions of key components within the proposed model, focusing on clinical data, the CEA module, the SMA module, and dynamic weight (DW) management. To evaluate the effectiveness of these components, we developed two alternative approaches: 1) excluding the CEA or SMA module, where feature fusion is done through simple data concatenation; 2) excluding the dynamic weight module, where loss weights for classification and segmentation are fixed at 0.5/0.5. These alternatives were trained alongside our original method, maintaining consistency in all other training parameters. We then evaluated each model's performance on the same test dataset using standard metrics. Detailed results are presented in Table II.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

Initially, the baseline model, which relied solely on clinical data, achieved a classification Acc of 63.97%. Adding CT data increased the Acc to 70.56%, a 6.59% improvement. However, this integration led to a decline in the Dice coefficient for segmentation, dropping from 82.27% to 78.33%, suggesting that handling multiple tasks might compromise segmentation performance. Introducing the CEA module, which effectively integrates multimodal data, further improved classification Acc to 89.71%. Adding the SMA module, which leverages hidden segmentation features to enhance classification, maintained strong classification performance at 89.34%, while segmentation Dice improved to 81.44%.

The implementation of dynamic weight management significantly enhanced both classification and segmentation tasks. Classification Acc rose to 91.91%, marking a 2.57% improvement over the previous setup, while the Dice coefficient reached 83.98% and IoU increased to 72.12%, surpassing the performance when focusing solely on segmentation. These findings underscore the importance of each component. The CEA module initiates substantial improvements, the SMA module maintains robust classification through segmentation features, and dynamic weight management effectively balances the two tasks, maximizing the model's overall performance.

2) Ablation Study on Hierarchical Feature Combinations: This ablation study evaluates the influence of hidden segmentation features from various Transformer layers ($\mathbf{Z}_3$, $\mathbf{Z}_6$, $\mathbf{Z}_9$, and $\mathbf{Z}_{12}$) on segmentation-assisted classification. As detailed in Table III, leveraging only the deepest feature, $\mathbf{Z}_{12}$, yields optimal performance across all metrics, highlighting that deeper features provide the most discriminative semantic information for both segmentation and classification tasks.

When assessing individual shallow features, clear hierarchical differences emerge: $\mathbf{Z}_9$ outperforms $\mathbf{Z}_6$, while $\mathbf{Z}_3$ yields the weakest results. This progression demonstrates that feature representational capacity increases with network depth, as shallow features primarily encode low-level spatial information. In combinations of multiple features, configurations containing $\mathbf{Z}_{12}$ consistently deliver strong results, whereas those excluding it exhibit noticeable performance degradation. The combination of $\mathbf{Z}_9$ and $\mathbf{Z}_{12}$ shows effective complementarity, enhancing overall performance. However, combining all four features results in lower performance compared to the optimal single-feature setup, suggesting that excessive feature stacking introduces redundancy and noise.

In summary, the study underscores the critical role of deep features, such as $\mathbf{Z}_{12}$, in segmentation-assisted classification. It emphasizes the importance of strategic feature selection over indiscriminate stacking to achieve optimal multimodal integration and maximize overall performance.

3) Ablation Study on Loss Components: To further investigate the impact of different loss functions on the performance of segmentation-assisted classification, we conducted an ablation study with three combinations: $\mathcal{L}_{\text{BCE}}+\mathcal{L}_{\text{Dice}}$, $\mathcal{L}_{\text{Focal}}+\mathcal{L}_{\text{Dice}}$, and $\mathcal{L}_{\text{BCE}}+\mathcal{L}_{\text{Focal}}+\mathcal{L}_{\text{Dice}}$, as shown in Table IV. When a loss term was absent, its weight was reallocated to the remaining classification loss to ensure fair comparisons.

The results demonstrate that the full combination of $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{Focal}}$, and $\mathcal{L}_{\text{Dice}}$ yields the best performance across both classification and segmentation tasks. Specifically, $\mathcal{L}_{\text{BCE}}$ provides stable gradient signals, $\mathcal{L}_{\text{Focal}}$ focuses on hard-to-classify examples, and $\mathcal{L}_{\text{Dice}}$ ensures accurate region-level supervision. In contrast, the two-loss combinations show inferior performance, suggesting that missing any one component weakens the model's ability to optimize both tasks jointly. These findings highlight that a well-balanced integration of complementary loss functions is critical for achieving robust and comprehensive performance in multitask learning.

F. Comparative Study of Adaptive Loss Functions

1) Performance Comparison: Fixed and Adaptive Methods: This study explores the advantages of adaptive loss functions and compares their effectiveness across various tasks. Through a series of experiments, we assessed how adaptive loss functions can enhance model performance compared to fixed-weight configurations, with the results detailed in Table V.

Different fixed-weight settings had varying impacts on performance. For example, a weight ratio of 0.5/0.5 yielded a classification Acc of 89.34%, an F1 of 86.64%, and a Dice coefficient of 81.44%, showing competitive performance. This suggests that while a balanced fixed-weight approach can sustain strong performance, it lacks the nuanced adaptability that automatic adjustment methods offer. Other configurations, such as weight ratios of 0.7/0.3 and 0.9/0.1, prioritized classification Acc, achieving 89.71% and 90.44%, respectively. However, this emphasis on classification led to compromises in segmentation quality, with HD95 values increasing to 4.18mm and 4.78mm, illustrating the trade-off between classification performance and segmentation accuracy when segmentation is deprioritized.

These findings highlight the flexibility of adaptive loss functions, which dynamically adjust task-specific weight distributions, thereby overcoming the limitations of fixed-weight settings. By implementing adaptive loss functions with automatic weight adjustment, we observed significant improvements in both classification and segmentation tasks. Classification Acc reached 91.91%, and the segmentation performance achieved a Dice coefficient of 83.98%, IoU of 72.12%, and HD95 of 3.73mm. This substantial enhancement underscores the efficacy of the adaptive approach in achieving optimal performance across tasks, demonstrating its ability to adjust loss weights according to task-specific needs, leading to superior multi-task learning outcomes.

2) Empirical Validation of Dynamic Loss Threshold: This study examines the impact of threshold selection for the dynamic loss function through sensitivity analysis. As illustrated in Figure 5, classification accuracy was evaluated across various Dice thresholds ranging from 0.7 to 0.9. The analysis revealed that a threshold of 0.8 achieved the highest accuracy of 91.91%, outperforming all other values. Thresholds below 0.78 resulted in decreased accuracy, while those above 0.8 showed no additional improvement.

These results confirm that 0.8 is the optimal threshold for balancing segmentation and classification tasks within the multi-task framework. By ensuring a smooth transition be-

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

tween segmentation-focused and classification-focused training, the selected threshold contributes significantly to the overall performance enhancements demonstrated in this study.

G. Discussions and Limitations

1) Advantages and Innovations: The USCNet framework makes three fundamental contributions to multimodal medical image analysis. First, its unified architecture combining CT imaging and EHR data establishes a new paradigm for segmentation-assisted classification in Urolithiasis management. Second, the novel cross-attention mechanism provides interpretable fusion of radiographic and clinical features through learnable projection matrices, outperforming traditional concatenation approaches. Third, the dynamic loss function implements curriculum learning by automatically shifting focus from segmentation to classification tasks during training, achieving superior optimization balance compared to fixed-weight schemes.

In addition to its technical advantages, USCNet offers distinct clinical benefits. Traditional methods require postoperative specimen images as input, limiting their practical clinical use. Conversely, our model innovatively utilizes preoperative imaging and clinical data to anticipate the infectivity of stones, providing essential support for decision-making during perioperative management and potentially minimizing postoperative complications. The predictive outcomes play a pivotal role in guiding clinical decision-making. For infectious stones, the model enables clinicians to initiate targeted antibiotic therapy prior to surgery and refine surgical strategies accordingly. Conversely, for non-infectious stones, it helps avoid unnecessary antibiotic use, allowing a greater focus on metabolic management and tailored treatment approaches. Importantly, USCNet is accessible not only to well-equipped medical institutions but also to those with limited resources. This adaptability allows physicians at various experience levels to make accurate diagnoses, thereby enhancing equitable access to high-quality urological care.

2) Limitations and Future Directions: While USCNet offers several advantages, there are notable limitations that must be addressed. First, the model's performance heavily relies on the completeness and quality of multimodal data, particularly the accuracy and availability of laboratory test results in EHR. Second, its generalization capability requires improvement in handling complex clinical scenarios, such as cases with anatomical variations or multiple stone types. Third, models developed using single-center data are prone to spectrum bias, which may limit their applicability to datasets from other medical institutions. Fourth, the computational complexity of the Transformer architecture constrains its use in resource-limited environments. Future research will aim to address these challenges by focusing on the development of lightweight network architectures to enhance computational efficiency. Additionally, multi-center validation studies will be conducted to better assess and improve the model's generalization capabilities. Efforts will also be directed toward expanding clinical applications, including real-time stone analysis and surgical navigation, to broaden the practical utility of the model.

IV. Conclusion

This study proposed a novel multi-task USCNet model for precise segmentation and classification of urinary stones in preoperative settings. This method achieved effective integration of CT images and electronic health records through a dedicated CT-EHR attention module, while employing an SMA module that significantly improves feature extraction accuracy compared to conventional approaches. The introduced dynamic loss function effectively balanced multi-task learning with accelerated convergence. Experimental results demonstrated superior performance compared to existing methods on our proprietary kidney stone dataset across multiple classification metrics. However, the model exhibited certain limitations, including a strong dependence on the quality of multimodal data, limited generalizability in complex clinical scenarios, and relatively high computational complexity. Future research will focus on designing a lightweight architecture for computational efficiency and extending the model's applications to broader clinical settings, including real-time stone analysis and surgical planning assistance.

V. Compliance With Ethical Standards

This retrospective study was approved by the institutional review board of Longgang District People's Hospital of Shenzhen (2022065).

References

[1] X. Qian, J. Wan, J. Xu, C. Liu, M. Zhong, J. Zhang, Y. Zhang, and S. Wang, "Epidemiological trends of urolithiasis at the global, regional, and national levels: a population-based study," *International Journal of Clinical Practice*, Mar. 2022.

[2] X. W. Lin, Y. H. Luo, J. L. Li, and B. Y. Zhang, "Epidemiological research progress on urological stones and stone composition," *Asian Journal of Clinical Medicine*, vol. 4, no. 2, p. 49, Mar. 2021.

[3] W. Wang, J. Fan, G. Huang, J. Li, X. Zhu, Y. Tian, and L. Su, "Prevalence of kidney stones in mainland china: A systematic review," *Scientific Reports*, vol. 7, no. 1, p. 41630, Jan. 2017.

[4] A. Alpendri and H. R. Danarto, "Stone composition pattern of kidney stone," *Jurnal Urologi Indonesia*, vol. 20, no. 1, p. 44, Jan. 2013.

[5] B. Yang, D. Wang, Y. Zhou, G. Zhou, C. Wan, J. Xu, and J. Liu, "Research advances of ct and ai technology in predicting the composition of urinary calculi," *Journal of Clinical Urology*, vol. 38, no. 2, pp. 139–145, Feb. 2023.

[6] C. Türk, A. Petřík, K. Sarica, C. Seitz, A. Skolarikos, M. Straub, and T. Knoll, "Eau guidelines on diagnosis and conservative management of urolithiasis," *European Urology*, vol. 69, no. 3, pp. 468–474, Mar. 2016.

[7] M. S. Pearle, D. S. Goldfarb, D. G. Assimos, G. Curhan, C. J. Denu-Ciocca, B. R. Matlaga, M. Monga, K. L. Penniston, G. M. Preminger, T. M. Turk *et al.*, "Medical management of kidney stones: Aua guideline," *Journal of Urology*, vol. 192, no. 2, pp. 316–324, Aug. 2014.

[8] S. Gayathri, J. Gowthami, S. Jayavarshini, K. Karthika, and K. Nandhini, "Kidney stone prediction based on urine analysis using ensemble learning," *2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0*, Apr. 2025.

[9] A. Basiri, M. Taheri, and F. Taheri, "What is the state of the stone analysis techniques in urolithiasis?" *Urology Journal*, vol. 9, no. 2, pp. 445–454, May 2012.

[10] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps: Automation of Decision Making*, pp. 323–350, Nov. 2017.

[11] A. R. Kambadakone, B. H. Eisner, O. A. Catalano, and D. V. Sahani, "New and evolving concepts in the imaging and management of urolithiasis: urologists' perspective," *Radiographics*, vol. 30, no. 3, pp. 603–623, May 2010.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3680589

[12] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision transformers, ensemble model, and transfer learning leveraging explainable ai for brain tumor detection and classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1261–1272, 2024.

[13] M. Y. Sayed, Y. Waykar, B. Nepal, S. Y. Pabalkar, M. Vaishnavi, and P. K. Sudhakar, "Hybrid neural network framework for multiclass classification of kidney stones from ct scans," *2025 6th International Conference for Emerging Technology (INCET)*, May. 2025.

[14] W. Xu, C. Lai, Z. Mo, C. Liu, M. Li, G. Zhao, and K. Xu, "Clinical-inspired framework for automatic kidney stone recognition and analysis on transverse ct images," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[15] Z. H. A. N. Xiaofei, C. U. I. Zhenyu, and Z. H. A. O. Chunli, "Advances on artificial intelligence in the diagnosis and treatment of urinary calculi," *Journal of Hebei Medical College for Continuing Education*, vol. 39, no. 4, p. 30, 2022.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] J. Hines, J. A. Karajgikar, J. D. Giardina, and B. Friedman, *Imaging the Urinary Tract: Fundamentals of Ultrasound, Computed Tomography, and Magnetic Resonance Imaging*. Interventional Urology, Nov. 2021.

[18] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, Mar. 2017.

[19] S. Asif, M. Zhao, X. Chen, and Y. Zhu, "Stonenet: An efficient lightweight model based on depthwise separable convolutions for kidney stone detection from ct images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 15, no. 4, pp. 633–652, Jul. 2023.

[20] D. C. Elton, E. B. Turkbey, P. J. Pickhardt, and R. M. Summers, "A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast ct scans," *Medical Physics*, vol. 49, no. 4, pp. 2545–2554, Feb. 2022.

[21] K. K. Patro, J. P. Allam, B. C. Neelapu, R. Tadeusiewicz, U. R. Acharya, M. Hammad, O. Yildirim, and P. Plawiak, "Application of kronecker convolutions in deep learning technique for automated detection of kidney stones with coronal ct images," *Information Sciences*, vol. 640, p. 119005, Sep. 2023.

[22] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, and H. Lu, "Resganet: Residual group attention network for medical image classification and segmentation," *Medical Image Analysis (MIA)*, vol. 76, p. 102313, Feb. 2022.

[23] M. Baharoon, H. Almatar, R. Alduhayan, T. Aldebasi, B. Alahmadi, Y. Bokhari, M. Alawad, A. Almazroa, and A. Aljouie, "Hymnet: a multimodal deep learning system for hypertension classification using fundus photographs and cardiometabolic risk factors," *arXiv preprint arXiv:2310.01099*, Mar. 2024.

[24] X. Yu, A. Elazab, R. Ge, J. Zhu, L. Zhang, G. Jia, Q. Wu, X. Wan, L. Li, and C. Wang, "Ich-prnet: a cross-modal intracerebral haemorrhage prognostic prediction method using joint-attention interaction mechanism," *Neural Networks*, vol. 184, p. 107096, Apr. 2025.

[25] X. Yu, A. Elazab, R. Ge, H. Jin, X. Jiang, G. Jia, Q. Wu, Q. Shi, and C. Wang, "Ich-scnet: Intracerebral hemorrhage segmentation and prognosis classification network using clip-guided sam mechanism," in *2024 IEEE International Conference on Bioinformatics and Biomedicine)*. IEEE, Dec. 2024, pp. 2795–2800.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2023.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Jun. 2020.

[28] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Nov. 2015, pp. 234–241.

[30] Z. Xing, L. Zhu, L. Yu, Z. Xing, and L. Wan, "Hybrid masked image modeling for 3d medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 2115–2125, 2024.

[31] W. Zhu, R. Zhou, Y. Yao, T. D. Campbell, R. K. Jain, and J. Luo, "Segprompt: Using segmentation map as a better prompt to finetune deep models for kidney stone classification," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1680–1690.

[32] N. Saeed, I. Sobirov, R. Al Majzoub, and M. Yaqub, "Tmss: an end-to-end transformer-based multimodal network for segmentation and survival prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Nature Switzerland, Sep. 2022, pp. 319–329.

[33] G. Wang, X. Lou, F. Guo, D. Kwok, and C. Cao, "Ehr-hgcn: An enhanced hybrid approach for text classification using heterogeneous graph convolutional networks in electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1668–1679, 2023.

[34] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[35] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[36] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2021.

Authorized licensed use limited to: Chiba University. Downloaded on May 19,2026 at 07:13:51 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.