IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 May 2025, accepted 28 May 2025, date of publication 9 June 2025, date of current version 18 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3577959

RESEARCH ARTICLE

An Ensemble Deep Learning Approach for Accurate Urinary Sediment Detection Using YOLOv9e and KD-YOLOX-ViT

MANSURA NAZNINE[1,2], ABDUS SALAM[1,3], MUHAMMAD MOHSIN KHAN[4], SADIA FARHANA NOBI[1,5], AND MUHAMMAD E. H. CHOWDHURY[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Qatar University, Doha, Qatar
[2]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh
[3]Department of Electrical and Computer Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh
[4]Department of Neurology, Hamad Medical Corporation, Doha, Qatar
[5]Department of Public Health and Informatics, BSMMU, Dhaka 1205, Bangladesh

Corresponding author: Muhammad E. H. Chowdhury (mchowdhury@qu.edu.qa)

The open access publication of this article is supported by Qatar National Library.

ABSTRACT Urinary Tract Infections (UTIs) are common medical diseases that occur from bacterial infections which can impact any area of the urinary system. Detecting urine particles is crucial for diagnosing UTIs and other renal abnormalities, as the presence of red blood cells (RBCs), white blood cells (WBCs), casts, fungi, and bacteria in urine can indicate disorders such as hematuria, kidney stones, and urinary tract malignancies. This study introduces an advanced deep learning method that utilizes ensemble of YOLOv9e and KD-YOLOX-ViT models. YOLOv9e model is built upon the Generalized Efficient Layer Aggregation Network (GELAN) and Programmable Gradient Information (PGI). The KD-YOLOX-ViT model integrates knowledge distillation and Vision Transformer (ViT) modules within the YOLOX framework to improve performance. This ensemble framework combines the strengths of YOLOv9e and KD-YOLOX-ViT models to automate the detection and classification of seven distinct urine sediment particles, effectively addressing challenges related to class imbalance, image resolution, and domain adaptation. The YOLOv9e model exhibited exceptional performance, obtaining precision, recall, and mean average precision (mAP50) scores of 88.5%, 88.1%, and 92.2%, respectively. The KD-YOLOX-ViT model also performed well, achieving precision, recall, mAP50, and mAP50-95 scores of 86%, 88%, 86.7%, and 53.3%, respectively. By ensembling YOLOv9e and KD-YOLOX-ViT using Weighted Box Fusion (WBF), the model achieved a final mAP50 of 94.18%. The model's adaptability was further validated through rigorous external validation on a novel dataset, yielding in a mAP50 of 94.64%. Comparative analysis with state-of-the-art methods confirms the model's real-time analytical capabilities. Moreover, the integration of eXplainable AI (XAI) enhances interpretability, offering valuable insights for confident clinical diagnosis. This comprehensive approach shows significant promise in advancing diagnostic accuracy and enabling earlier, real-time treatment on a global scale.

INDEX TERMS Urine sediment detection, YOLOv9e, KD-YOLOX-ViT, external validation, ensembling methods.

I. INTRODUCTION

Urinary Tract Infections (UTIs) and other renal abnormalities are prevalent medical conditions that can have a substantial influence on any section of the urinary system, such as the kidneys, ureters, bladder, and urethra. Urine tract infections (UTIs) commonly happen when bacteria, usually from the gastrointestinal system, enter the urine tract through the urethra and reproduce, causing inflammation and irritation of the tissues in the urinary tract [1]. The presence of germs in the body leads to symptoms such as discomfort, a sensation of burning during urination, frequent urine, and a strong

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

need to urinate. Without prompt medical attention, the infection might progress to the kidneys, resulting in more severe symptoms and possible consequences. Renal abnormalities, such as the detection of red blood cells (RBC), white blood cells (WBC), casts, fungus, and bacteria in urine, can serve as indicators for a range of disorders including hematuria, kidney stones, and urinary tract malignancies [2]. Both urinary tract infections (UTIs) and other kidney abnormalities have a substantial influence on worldwide health, resulting in approximately 830,000 deaths and 18,467,000 disability-adjusted life years each year. Obtaining a comprehensive understanding of the underlying causes behind these disorders is essential in order to create efficient diagnostic and therapeutic approaches for managing and preventing these infections and abnormalities [3].

The identification of renal abnormalities is dependent on a combination of clinical evaluation, urine dipstick analysis, urine culture, and thorough urinalysis. At first, healthcare personnel assess lower urinary tract symptoms (LUTS) such as dysuria, polyuria, hematuria, and pelvic pain [4]. Urine dipstick testing is a quick and cost-efficient way to identify infection-related compounds such as leukocytes, nitrites, and blood [5]. In order to obtain a more conclusive diagnosis, urine cultures are performed to identify the precise bacteria responsible for the infection and establish the most suitable antibiotic treatment. In addition, a thorough urinalysis, which includes a microscopic examination, offers extensive information about the presence of casts, crystals, epithelial cells, red blood cells, white blood cells, and fungus [6]. This helps in evaluating the severity and specific features of urinary tract infections and other kidney disorders. RBCs in the urine can be a sign of hematuria and bladder or urinary tract malignancy, whereas the presence of crystals suggests the existence of kidney or urinary tract stones. White blood cells (WBCs) and epithelial cells are useful in the diagnosis of urinary tract infections (UTIs), and the presence of epithelial cells in urine may also suggest the presence of specific forms of cancer [7].

The existing diagnostic techniques for urinary tract infections (UTIs) and kidney abnormalities have numerous notable constraints, particularly when used in point-of-care settings [8]. Urine dipstick tests and clinical assessments frequently lack the required sensitivity and specificity to accurately detect urinary tract disorders. Urine sediment analysis is an important technique for identifying kidney problems, but it involves a laborious and time-consuming process [9]. This approach includes centrifuging fresh urine samples and doing a microscopic examination by highly skilled specialists [10]. The conventional approach usually requires a time frame of 24 to 48 hours, which turns it inappropriate for prompt diagnosis and fast clinical decision-making. Urine microscopy is limited to well-equipped laboratories due to the need for trained workers and specialized equipment, making it unavailable in many primary care and urgent care settings. Moreover, the process of manually examining urine sediment is susceptible to human variability, subjective analysis, and reliance on the operator, especially in laboratories that handle a large number of samples. The limitations emphasize the immediate requirement for more effective, automated, and easily available diagnostic techniques to enhance the identification and treatment of kidney disorders [11].

Recent research has utilized deep learning (DL) [12], [13], [14], [15], [16] approaches to improve both the accuracy and effectiveness of automated urine sediment analysis. These sophisticated models have shown major potential in identifying and categorizing urine particles, simplifying diagnostic procedures in clinical environments. Nevertheless, despite these advances, numerous limitations remain, signifying a demand for additional investigation. Current models often struggle with accurately distinguishing particle subtypes, leading to potential diagnostic inaccuracies [11], [17]. Additionally, the reliance on limited datasets raises concerns about the generalizability of these models across diverse clinical scenarios. While some methods achieve high accuracy, they frequently lack thorough testing across varied datasets, limiting their real-world applicability. This research seeks to address these challenges by developing a more robust and generalizable approach to urine sediment detection. The main contributions of this study are stated below:

1. The YOLOv9e and KD-YOLOX-ViT models were successfully utilized for the precise detection and classification of urine sediment particles, demonstrating their individual strengths and suitability for medical image analysis. This study evaluates their accuracy and highlights their potential applicability in the domain of automated urine sediment analysis.
2. Additional dataset has been used for extensive external validation, enhancing the validity of the results and addressing issues regarding the generalizability of existing approaches. The model's constant performance in multiple datasets demonstrates its ability to adapt in a wide range of clinical settings.
3. An ensemble framework combining YOLOv9e and KD-YOLOX-ViT was proposed, utilizing Weighted Box Fusion (WBF) to aggregate predictions from both models. This integration of complementary architectures led to enhanced detection accuracy and robustness, significantly improving the overall efficacy and reliability of urine sediment analysis. This approach marks a notable advancement in achieving higher precision in this domain.

The article is divided into five sections. **Section I** provides a brief explanation of the study's motivation, as well as the challenges involved in detecting urine sediment. **Section II** will discuss similar works and their contributions and limitations. In **Section III**, we provide a conceptual framework that serves as the foundation of the proposed methodology. The results of this research and the analysis of its constraints and potential future developments are concisely presented in **Section IV**. **Section V** includes conclusion.

II. RELATED WORKS

The incorporation of deep learning [12], [13], [17], [18], [19] models in medical diagnostics has greatly progressed the field of automated urine sediment analysis. These models are increasingly used to address the limits of manual microscopy, providing improved accuracy, speed, and objectivity [20], [21], [22], [23], [24]. This literature review examines the present condition of deep learning applications in urine sediment detection, with a specific emphasis on the progress made, difficulties observed and promising possibilities for future investigation.

Lyu et al. [17] presented YUS-Net, an innovative deep learning model designed to automatically detect urine sediment particles using the YOLOX framework. This approach combines attention processes and a tailored data augmentation strategy to improve the process of extracting features from microscopic urine photos. YUS-Net effectively tackles the issue of class imbalance by utilizing Varifocal loss, resulting in a mean average precision (mAP) of 96.07% on the USE dataset. This study emphasizes the effectiveness of the approach in identifying difficult particle types, such as casts and epithelial cells. Nevertheless, YUS-Net has many shortcomings such as its incapability to differentiate between different particle subtypes and its dependence on a single dataset, which gives rise to questions over its applicability to other scenarios. The research conducted by Suhail and Brindha [19] assessed the effectiveness of an identification system based on EGA-YOLOv5. Among the many versions of YOLOv5, it was discovered that YOLOv5l had the greatest mean average precision (mAP) of 85.8%. It particularly excelled in accurately identifying erythrocytes and casts. This study also emphasizes the superior speed and accuracy of YOLOv5 models in comparison to conventional CNN-based approaches. Nevertheless, the model's dependence on a restricted dataset and its struggle to differentiate between subclasses highlight the necessity for larger datasets and more varied testing conditions in order to enhance the accuracy of detection and its capacity to apply to different scenarios. Liang et al. [11] improved urine sediment detection by utilizing a DenseNet-enhanced Feature Pyramid Network (DFPN), resulting in notable contributions. Their study was centered around tackling problems such as class ambiguity and cell adhesion in USE images. By including attention mechanisms and fine-tuning on the COCO dataset, they achieved a mean average precision (mAP) of 86.9%. This approach notably enhanced the precision of erythrocyte detectio. Although DFPN's capability to extract detailed characteristics is remarkable, its performance may differ depending on the diversity of the dataset. This indicates that additional validation on diverse datasets is required to verify the reliability of DFPN in clinical applications.

In spite of the promising progress, there are still some obstacles that need to be addressed. The research conducted by Avci et al. [16] demonstrates the constraints associated with depending on artificial datasets to train deep learning models. Their incorporation of super-resolution techniques into Faster R-CNN yielded impressive classification accuracies. However, the absence of real-world data during the training phase raises doubts about the model's suitability for various clinical contexts. Li et al.'s [15] implementation of RetinaNet with ResNet and FPN yielded an accuracy of 88.65%. However, they observed difficulties related to misclassification and the presence of overlapping particles. The performance of the model may fluctuate depending on the quality of the samples, the techniques used for preparation, and the presence of artifacts. This highlights the importance of doing more rigorous validation in various clinical contexts. Chan et al. [14] tackled the problem of limited annotated data in medical imaging by proposing a few-shot object recognition method specifically designed for urine sediment analysis. Their method achieved a notable enhancement in detection accuracy for new classes by incorporating Background Suppression Attention (BSA) and Feature Space Fine-tuning (FSF) modules into a Faster R-CNN framework. Nevertheless, the study's dependence on few-shot learning approaches emphasizes the persistent difficulty of limited data availability in the field of medical imaging. Although this technique shows potential for enhancing diagnostic accuracy, additional study is required to guarantee that these models can be successfully used in a wider variety of clinical settings.

Despite the fact that deep learning models have demonstrated significant potential in enhancing automated urine sediment analysis, there are still numerous existing constraints. The use of limited and occasionally artificial datasets gives rise to questions over the ability of these models to be used effectively and consistently in a wide range of clinical situations. Moreover, the present models have a substantial hurdle in failing to differentiate between different types of particles and effectively manage complicated scenarios where particles overlap. Further studies should prioritize the expansion of datasets, development of model generalization, and incorporation of more advanced post-processing approaches that improve the accuracy and feasibility of these models in real-world hospital environments.

III. EXPERIMENTAL METHODOLOGY

This study presents a framework for multiclass object detection of urine sediment particles using the Urine Sediment Dataset (USE) and an independent clinical microscopy dataset for external validation. After preprocessing, including annotation conversion and bounding box extraction, several YOLO-based models were trained and evaluated. An ensemble approach combining YOLOv9e and KD-YOLOX-ViT was used, with predictions aggregated via Weighted Box Fusion (WBF). EigenCAM was applied for model interpretability. The framework demonstrated high performance and generalizability, highlighting its potential for clinical diagnostic applications.

A. DATASET DESCRIPTION

The Urine Sediment Dataset (USE) [11], [13], [25] utilized in this study is derived from a publicly accessible source,

TABLE 1. Dataset details.

The Urine Sediment Dataset (USE)	
Class	**Instances**
Erythrocytes (eryth)	21,815
Leukocytes (leuko)	6,169
Epithelial cells (epith)	6,175
Crystals (cryst)	1,644
Casts	3,663
Mycetes	2,083
Epithelial nuclei (epithn)	687
A Clinical Microscopy Dataset	
Class	**Instances**
Rod	1697
RBC/WBC	1056
Yeast	41
Miscellaneous	550
Single EPC	182
Small EPC Sheet	26
Large EPC Sheet	10

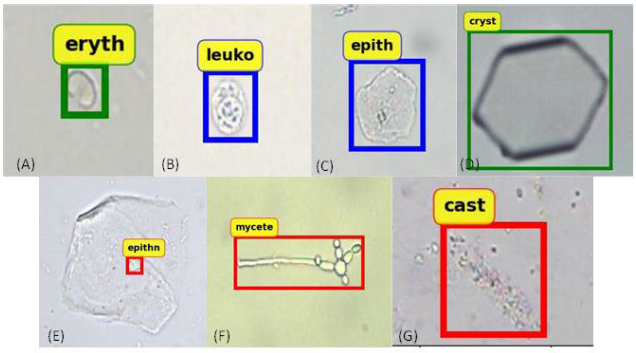

FIGURE 1. Urine sediment particles from USE dataset.

specifically curated for the detection of various components in urine sediment. This dataset is the largest urine sediment dataset available online, making it a valuable resource for research and development in the field. The USE dataset comprises a total of 5,376 images categorized into seven distinct cell types: casts, crystals (cryst), epithelial cells (epith), epithelial nuclei (epithn), erythrocytes (eryth), leukocytes (leuko), and mycetes. Notably, all types of casts are grouped into a single category. The dataset is divided into three subsets at random: test, validation, and training sets, in order to provide a thorough assessment. The training set constitutes 79% of the total images, amounting to 4,256 images. The test set consists of 16% of the images, or 852 images, whereas the validation set consists of 5% of the images, or 268 images. The dataset annotations are in Pascal VOC format. The distribution of instances per cell category within the dataset is presented at **Table 1**. **Figure 1** demonstrates various particles from USE dataset.

A clinical microscopy dataset [26] to develop a deep learning diagnostic test for urinary tract infection: Liou et al. have presented a dataset that includes 300 images and 3,562 manually annotated urinary cells. These cells are categorized into seven clinically significant classes. This dataset was enriched by the collection of unstained and untreated urine samples from symptomatic UTI patients at a specialist LUTS outpatient clinic in central London between April and August 2022. Each image was annotated using ilastik for binary semantic segmentation. The dataset is divided into three primary folders (img, bin_mask, mult_mask), each of which contains 300 files. It offers an exhaustive resource for advanced machine learning applications in UTI diagnosis, as it is the most recent dataset publicly available. This dataset has been employed in our experiment to facilitate external validation. The distribution of instances per cell category within the dataset is presented at **Table 1**.

B. OVERVIEW OF THE PROPOSED FRAMEWORK

To guarantee that the results of the detection of particles in urine sediment are both accurate and generalizable, several significant phases are incorporated into the proposed framework. **Figure 2** demonstrates the summary of the proposed framework. In order to improve the quality of the training data, the dataset is initially preprocessed. Next, a variety of cutting-edge YOLO (You Only Look Once) models are implemented. These models are chosen for their balance of speed and accuracy in object detection tasks. In order to enhance the detection performance, an ensemble approach is implemented. By combining the predictions of multiple YOLO models, the variance and bias inherent in individual models are reduced, leading to more robust and accurate results. Diverse features and learning patterns that may be overlooked by a single model are captured by this ensemble method. In order to illustrate the models' generalizability across various datasets, external validation is implemented. The trained models' effectiveness in real-world scenarios is evaluated by testing them on an independent dataset. Finally, EigenCAM (Class Activation Mapping) is employed to provide interpretability to the model's predictions. EigenCAM identifies and emphasizes the specific areas in the input images that have a major influence on the decision-making process of the model.

C. DATASET PREPROCESSING

Some crucial steps were implemented during the data preprocessing phase to prepare two distinct datasets for multiclass object detection tasks. At first, the annotations of the Urine Sediment Dataset (USE), which were initially formatted in Pascal VOC, were converted to YOLO format in order to comply with the specifications of the selected object detection models. The bounding box coordinates and class labels were mapped from the Pascal VOC XML format to YOLO format text files during this conversion procedure. Each annotation was methodically modified to incorporate normalized coordinates of the bounding boxes in regard to the image dimensions and class indices which were encoded in a specific manner. Simultaneously, an alternate procedure was implemented for the clinical microscopy dataset introduced by Liou et al. [26], which provided binary and multiclass segmentation masks instead of traditional bounding boxes.

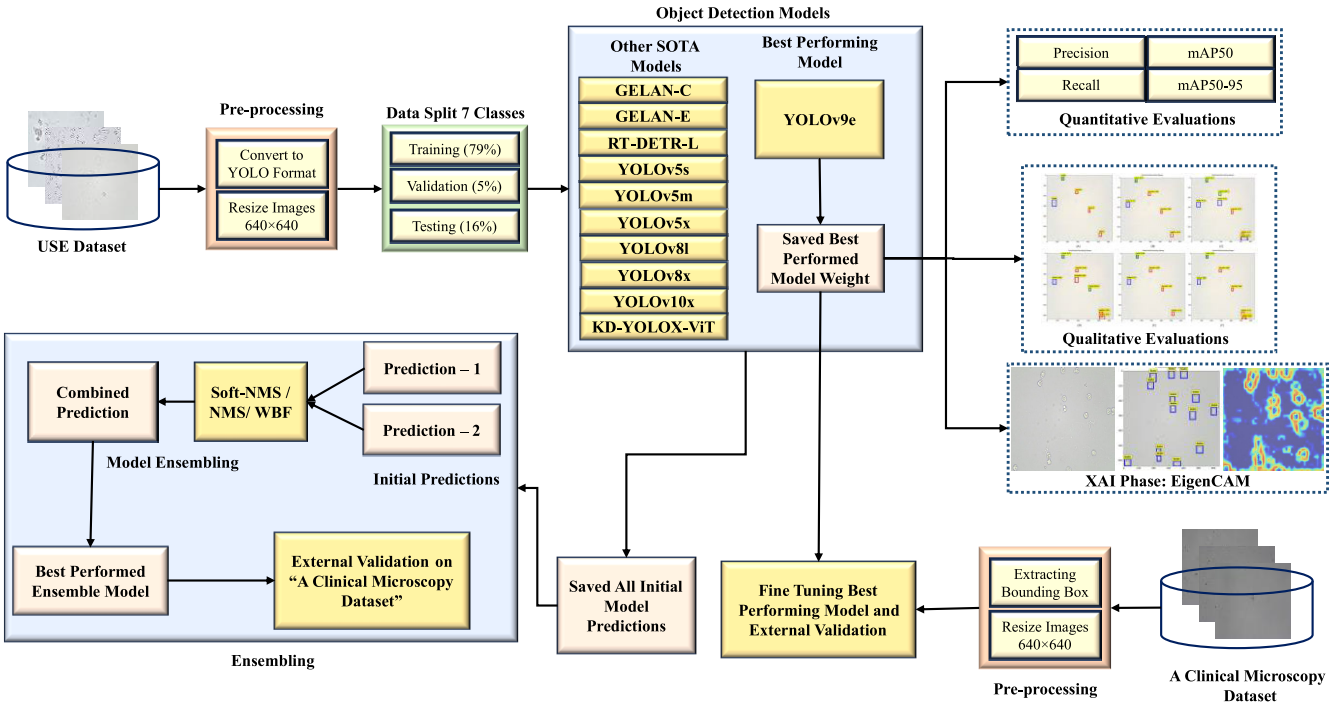

FIGURE 2. Summary of the proposed framework.

To facilitate multiclass object detection, this necessitated explicitly deriving bounding boxes from the segmentation masks. Furthermore, in order to guarantee optimal performance and consistency across both datasets, all images were uniformly resized to 640 × 640 pixels. The preprocessing methods were crucial for preparing the datasets for future training and evaluation, ensuring consistency with the selected methodology and objectives of the study.

In addition to the above preprocessing steps, several image enhancement techniques were explored to improve the quality of input data and potentially boost detection accuracy. Histogram Equalization was applied to enhance the global contrast of images by distributing pixel intensity values more uniformly. Contrast Limited Adaptive Histogram Equalization (CLAHE), a refined version of Histogram Equalization, was also tested to improve local contrast while limiting noise amplification. Similarly, Gamma Correction was employed to adjust image brightness and enhance features in both darker and lighter regions of the images. Although these techniques were implemented and their effects evaluated, they were not included in the final model due to a lack of significant improvement in performance. However, their inclusion in the preprocessing pipeline during initial experimentation underscored a thorough and systematic approach to optimizing the input data.

These preprocessing methods were essential for preparing the datasets for training and evaluation. They ensured consistency with the selected methodology, while the exploration of additional enhancement techniques, despite their exclusion from the final model, reflected a comprehensive effort to refine the data and enhance model performance.

D. EXPERIMENTAL DETAILS

Following the preprocessing phase, several deep learning-based object detection networks were trained on the prepared USE dataset to develop robust models. The networks utilized include YOLOv5, YOLOv8, YOLOv9, YOLOv10, RT-DETR and KD-YOLOX-ViT. By incorporating a mix of convolutional-based models like YOLO and transformer-enhanced models like KD-YOLOX-ViT, this experiment capitalizes on the strengths of both architectures.

1) YOLOV9

The YOLOv9e [27] model was selected for detecting urine sediments because of its advanced architecture along with the cutting-edge functionalities that effectively handle usual challenges in object detection using deep learning. YOLOv9, is one of the most recent versions in the YOLO series, represents a substantial advancement in real-time object detection through the incorporation of multiple key advancements that enhance accuracy, and efficiency.

The YOLO series improved object detection by analyzing full images in a single iteration through a convolutional neural network (CNN). This has gone through a series of improvements, progressing from YOLOv1 to YOLOv8, with each version including new ways to improve performance. YOLOv9 advances upon this foundation by handling the major issue of information bottleneck, which is common

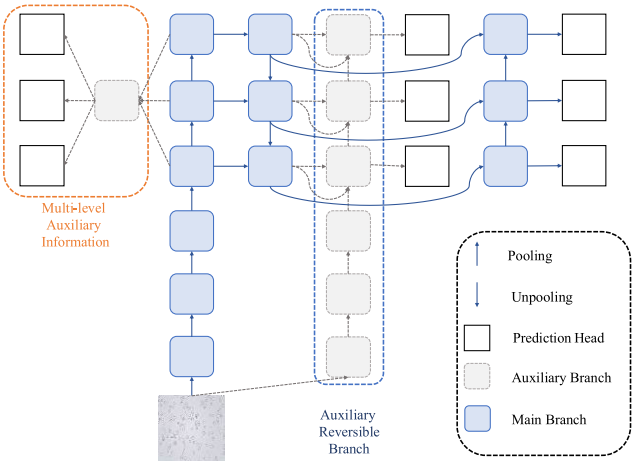

FIGURE 3. Programmable gradient information.

in deep neural networks as they become deeper and more complex.

a: PROGRAMMABLE GRADIENT INFORMATION (PGI)

YOLOv9-e incorporates Programmable Gradient Information (PGI) [28], an architecture specifically developed to address the information loss that naturally occurs during the propagation of data in deep neural networks. Architecture of PGI is presented in **Figure 3**. PGI enables the generation of dependable gradients by including an additional reversible branch alongside the primary inference pathway. The purpose of this auxiliary branch is to preserve the integrity of important data features, so ensuring that gradients become robust and informative during the whole training phase. YOLOv9e successfully resolves the problem of unstable gradients caused by information bottlenecks, especially in deep networks. The reversible branch enhances the network's capacity to acquire accurate correlations between inputs and targets, even in complicated settings, by providing additional gradient paths. The PGI framework is smoothly incorporated into the architecture of YOLOv9e, enabling the concurrent optimization of gradient information, parameter learning, and inference speed. By adopting this strategy, YOLOv9e is able to achieve exceptional performance while minimizing the computational expenses [29].

b: GENERALIZED EFFICIENT LAYER AGGREGATION NETWORK (GELAN-E)

In addition to PGI, the Generalized Efficient Layer Aggregation Network (GELAN-E) is an essential element of YOLOv9e. GELAN-E integrates the concepts of Spatial Pyramid Pooling (SPP) and Cross-Stage Partial Network (CSPNet) [30], along with an improved version of the Efficient Layer Aggregation Network (ELAN) [31]. GELAN-E incorporates advanced layer aggregation techniques that have been developed specifically to enhance feature extraction while preserving computational efficiency. This is accomplished by dividing the input features into several paths that undergo individual processing before they are combined again. The implementation of this dual-path technique not only accelerates the flow of gradients but also improves the network's capacity to capture and reuse features across several layers. GELAN-E's architecture incorporates spatial pooling operations that collect contextual information at many scales, enabling YOLOv9-e to accurately identify objects of different sizes and complexities. GELAN-E combines the gradient efficiency of CSPNet with the speed-focused design of ELAN, enabling YOLOv9e to achieve a balance between depth and computational complexity. This results in a model that is both remarkably accurate and fast during inference [32]. **Figure 4** represents the structure of GELAN.

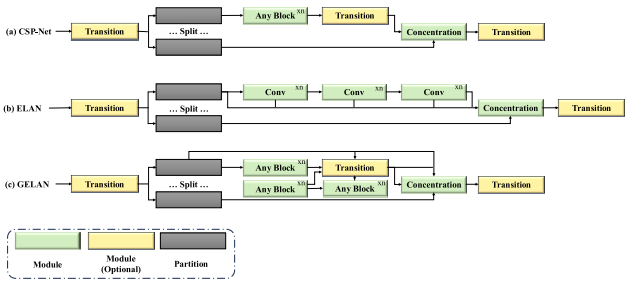

FIGURE 4. Structure of generalized efficient layer aggregation network (GELAN).

c: YOLOV9 ARCHITECTURE

The detailed architecture of YOLOv9 is shown in **Figure 5**. Building on previous YOLO versions, especially YOLOv7, YOLOv9 introduces several advancements to enhance feature extraction, computational efficiency, and detection accuracy [33]. The Backbone is central to YOLOv9, designed to balance extensive feature extraction and computational performance. It processes high-resolution images effectively without overloading resources. The Silence Block in the Backbone transfers input images to subsequent modules unchanged, ensuring seamless processing. Convolution Blocks, using 2D convolutions with Batch Normalization and SiLU activation, extract crucial features while preserving spatial dimensions through AutoPad. The RepNCSPELAN Block combines multi-scale feature extraction with computational efficiency, capturing intricate details critical for high-precision applications. The ADown Block facilitates down-sampling while preserving semantic and spatial details, and the SPPELAN Block employs Spatial Pyramid Pooling (SPP) for scale-invariant object representation.

The Neck refines features extracted by the Backbone for accurate object detection. An Up-Sampling Layer aligns feature resolutions, and the Concatenation Layer combines high- and low-resolution features for precise detection. Additional RepNCSPELAN Blocks [34] in the Neck enhance feature optimization for subsequent detection tasks.

The Auxiliary section improves training robustness and gradient flow, critical for model convergence and accuracy.

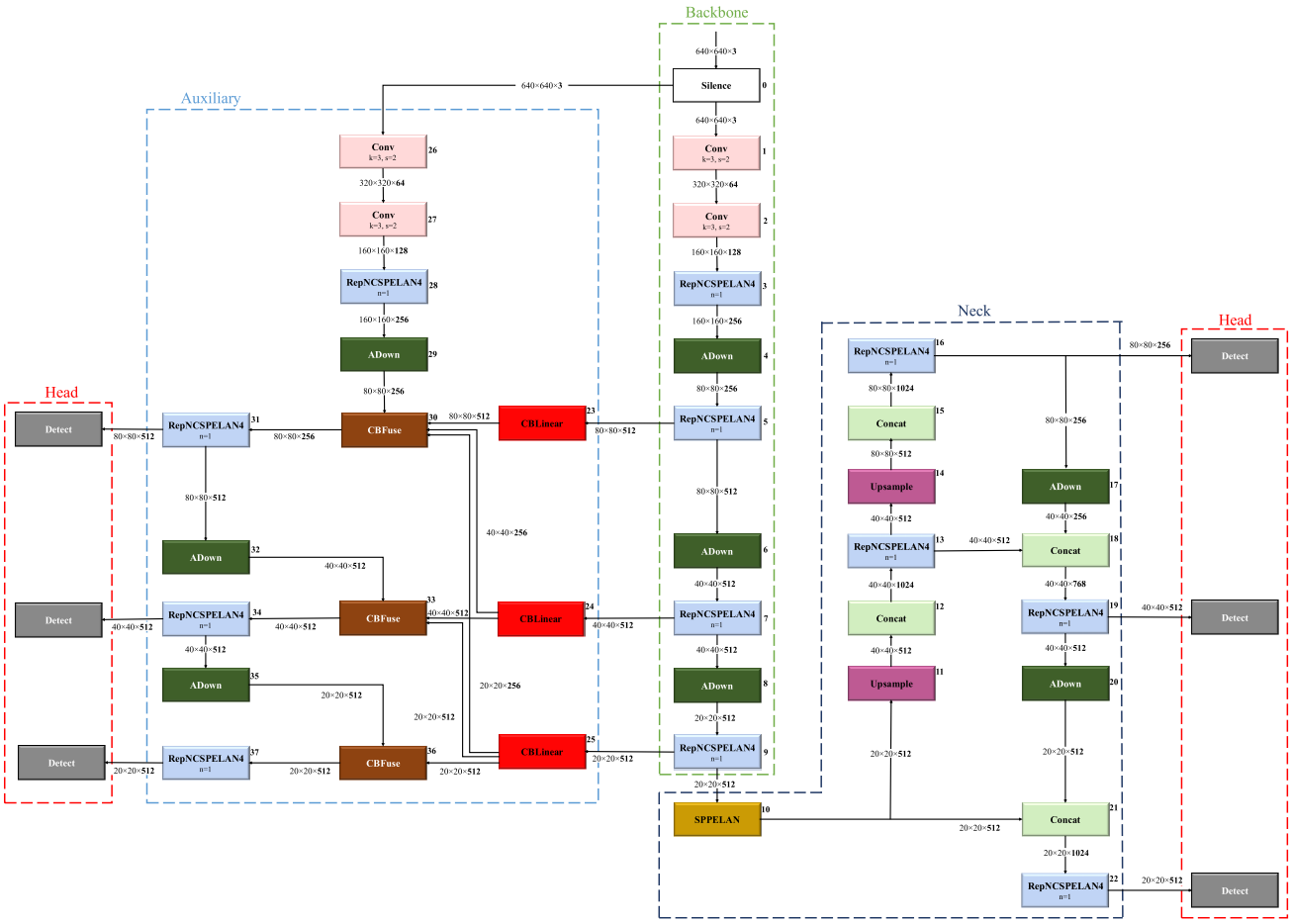

FIGURE 5. Architecture of YOLOv9.

CB-Linear Blocks generate feature maps at multiple scales, capturing high-level semantic details, while the CB-Fuse Block combines features from CB-Linear and ADown Blocks. This fusion enhances feature richness, improving interpretability and detection performance. The Auxiliary section operates only during training but plays a pivotal role in overall model optimization.

Finally, the Head of YOLOv9 executes the object detection tasks by taking the refined features from the Neck and producing the final predictions. The Detect Blocks in the Head are tailored for detecting objects at different scales—small, medium, and large—leveraging the refined features from the Neck to output precise bounding box coordinates and class predictions. This multi-scale approach ensures that the model can accurately detect objects regardless of their size within the image [35].

d: FINE-TUNING THE MODEL

The YOLOv9 model was fine-tuned to optimize its performance for the multiclass object detection of urine sediment types. The fine-tuning process involved selecting hyperparameters tailored to the dataset and task requirements. Stochastic Gradient Descent (SGD) was employed as the optimizer with an initial learning rate of 0.01 and momentum of 0.937 to ensure stable convergence.

The model was trained for 100 epochs with a batch size of 8, incorporating image augmentations such as mosaic (probability = 1.0), mixup (probability = 0.15), and horizontal flipping (probability = 0.5) to enhance generalization. Data augmentation parameters, including HSV (Hue = 0.015, Saturation = 0.7, Value = 0.4) and image translation ($\pm$0.1), were adjusted to simulate real-world variability in the dataset.

Bounding box regression and object detection thresholds were fine-tuned with an IoU training threshold of 0.2 and an anchor threshold of 5.0. Additionally, the warmup phase spanned 3 epochs with an initial momentum of 0.8 and a warmup learning rate of 0.1 for bias terms. These configurations aimed to stabilize early training and optimize convergence for effective multiclass detection.

This fine-tuning setup ensured the YOLOv9 model's robustness and adaptability for the given task, achieving a balance between precision and computational efficiency.

e: LOSS FUNCTION

The YOLOv9 model uses a composite loss function that integrates multiple components to optimize object detection

performance. These components include the box loss, class loss, and Distribution Focal Loss (DFL). Each component is weighted with specific gain factors to balance the contributions of the individual losses during training.

Box Loss: The box loss is responsible for minimizing the discrepancy between the predicted bounding box coordinates and the ground truth annotations. This loss component is crucial for ensuring accurate localization of objects within the image. The box loss is weighted by a factor of 7.5, signifying its strong contribution to the overall loss. A higher weight for box loss emphasizes the importance of precise bounding box predictions.

Class Loss: The classification loss ensures that the model correctly classifies the objects within the predicted bounding boxes. This loss is typically computed using Binary Cross-Entropy (BCE) loss for each object class, where the model is penalized for incorrect predictions and encouraged to assign higher confidence to the correct class. The classification loss is scaled by a factor of 0.5 to indicate its relative importance in comparison to the box loss, balancing the model's focus between object localization and classification.

Distribution Focal Loss (DFL): The Distribution Focal Loss (DFL) is introduced to refine bounding box regression by improving localization accuracy, especially in challenging cases where object boundaries may be difficult to predict. This loss focuses on the distribution of offsets for the bounding box, allowing the model to achieve better precision in its localization. The DFL is weighted by a factor of 1.5, emphasizing its role in fine-tuning bounding box predictions and enhancing detection accuracy.

2) KD-YOLOX-VIT: KNOWLEDGE DISTILLATION IN YOLOX-VIT

The YOLOX-ViT model [36], which integrates knowledge distillation (KD), represents a significant advancement in object detection, especially for systems with limited computational resources. By combining the YOLOX architecture with a Vision Transformer (ViT) layer [37], [38], the model improves detection accuracy while maintaining efficiency, especially in challenging environments such as underwater scenes, where traditional convolutional neural networks (CNNs) may not perform well. Knowledge distillation reduces the model's size without sacrificing performance by transferring knowledge from a larger teacher model (YOLOX-L) to a smaller student model (YOLOX-Nano-ViT). The distillation approach employs a combined loss function, consisting of both hard loss (ground truth matching) and soft loss (teacher-student output matching), ensuring that the student model captures essential details from the teacher.

The YOLOX-ViT architecture includes a backbone for feature extraction, a neck (FPN) for feature aggregation, and a decoupled head for bounding box regression and classification. The integration of the ViT layer between the backbone and neck enhances the model's ability to capture long-range dependencies, which is an improvement over the original YOLOX design. The knowledge distillation process unfolds in two stages: the training of the YOLOX-L teacher model, followed by the training of the YOLOX-Nano-ViT student model. During this phase, FPN outputs from both models are compared, and the soft loss is minimized. Additionally, by using an offline KD method, the training time is further reduced as inference results from the teacher model are stored and reused during the student's training, eliminating the need for real-time inference during each batch.

3) COMPARISON WITH OTHER STATE-OF-THE-ART MODELS

This study evaluates several state-of-the-art models to benchmark their efficiency in detecting and categorizing urine sediment particles, providing a comparative context for the proposed YOLOv9e model. RT-DETR-L [39], an advanced real-time detection model, improves upon the Detection Transformer (DETR) framework. Its transformer-based architecture effectively captures context and object relationships, making it well-suited for large-scale images and high-speed applications. Incorporating multi-scale feature extraction, dynamic head design, focal loss, and balanced feature alignment, RT-DETR-L demonstrates resilience and accuracy across diverse datasets and challenging scenarios.

The YOLOv5 series balances efficiency and accuracy, offering versions tailored to different computational constraints. YOLOv5s, the most lightweight, prioritizes speed for real-time applications at the expense of accuracy. YOLOv5m provides a middle ground, balancing speed and accuracy, while YOLOv5x, the largest version, optimizes detection for intricate or small objects. The series leverages CSP-Darknet [40], for efficient feature extraction, PANet [41] for effective feature propagation, and adaptive pooling [42], [43] to enhance localization accuracy.

YOLOv8 [44], developed by Ultralytics, represents a significant advancement over YOLOv3 and YOLOv5, excelling in object detection, segmentation, and classification tasks. It achieves improved mean Average Precision (mAP) with an efficient architecture that reduces parameters while enhancing performance. Notable features include the C2f block for enhanced feature representation, anchor-free object detection for reduced computational overhead, and upgraded 3×3 convolutions for efficiency. YOLOv8's sophisticated loss functions, such as CIoU and DFL, improve precision, particularly for smaller objects. Augmentation techniques during training further boost adaptability and accuracy.

E. EVALUATION MATRIX

Object detection models were evaluated using a range of performance parameters, including precision, recall, and mean Average Precision (mAP). These metrics are essential for measuring the accuracy of the models in recognizing and accurately locating objects in images. A standardized Intersection over Union (IoU) threshold of 0.5 to determine true positive identifications was employed.

$$Precision = \frac{TruePositive}{True\ Positive + False\ Positive} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

Precision is the percentage of positive predictions that are correctly identified out of all positive predictions made by the model. The proportion of correctly identified positive predictions out of all actual positive instances in the dataset is measured by recall.

Average Precision (AP) is calculated for a single class by evaluating the precision-recall (PR) curve, which plots precision against recall at different confidence thresholds. The area under this PR curve represents the AP for that class. Mathematically, the AP for a class c can be expressed as:

$$AP_c = \int_0^1 Precision\,(r)\,dr \quad (3)$$

where, $Precision\,(r)$ is the precision at recall r, and the integral accumulates the precision across different recall levels.

Mean Average Precision (mAP) aggregates the AP values across all classes to provide an overall performance measure. If AP_i denotes the Average Precision for class i and there are N classes, mAP is computed as:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (4)$$

mAP is frequently reported at a fixed IoU threshold, but performance can vary significantly with different thresholds. To address this, a more comprehensive evaluation involves calculating mAP at multiple IoU thresholds. The mAP50-95 metric averages the mAP values computed at various thresholds ranging from 0.5 to 0.95, with an increment of 0.05. This provides a more nuanced view of model performance across a spectrum of overlap criteria. Mathematically, mAP50-95 is defined as:

$$mAP_{50-95} = \frac{1}{10}\sum_{t=0.5}^{0.95} mAP_t \quad (5)$$

where, mAP_t represents the mean Average Precision calculated at IoU threshold t, and the summation averages these values over the specified range. This approach ensures a more robust evaluation by accounting for varying degrees of object overlap and offers a balanced assessment of detection accuracy.

F. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

EigenCAM [45] has played a crucial role in improving the interpretability of YOLOv9e model. It is essential for understanding the decision-making processes of deep learning models as they become increasingly common. These models, which are frequently intricate and undefined, necessitate the use of tools to make the outcomes understandable. EigenCAM, a technique developed by Muhammad et al. in 2020, is quite beneficial in this particular application. The method employs class activation maps (CAM) to visually represent the main components of features acquired by convolutional layers. The outcome is the creation of heatmaps superimposed on the original images, which emphasize the areas that have the greatest impact on the model's predictions. Unlike other systems like GradCAM [46], EigenCAM is highly valued for its simplicity, as it does not necessitate any re-training or changes. It conforms to human visual perception, enabling users to readily confirm whether the model's attention matches the human interpretation of significant visual components.

G. ENSEMBLE-BASED DETECTION REFINEMENT

A further investigation was conducted to assess the possibilities of merging predictions from the top-performing models in order to improve the accuracy of detection. A close study of the results from each model showed that their strengths often worked well together, suggesting that using more than one model together might produce better results. In order to do this, two ensemble techniques were utilized with the objective of enhancing the final predictions by utilizing the outputs of both models. The goal was to combine the projected bounding boxes from each model in a way that improves robustness and accuracy, therefore creating a more dependable detection method. **Figure 6** illustrates the block diagram of the ensembling method.

1) NON-MAXIMUM SUPPRESSION (NMS)

NMS is an approach used in object detection tasks to improve final predictions by removing redundant and overlapping bounding boxes. When used as an ensemble technique, NMS helps merge predictions from multiple models, ensuring that the final output consists of the most confident and non-overlapping detections. The NMS process begins with a list of detection boxes and their corresponding confidence scores. The algorithm first identifies the bounding box with the highest confidence score, represented as m. This box is then added to the final set of detections. Subsequently, the algorithm evaluates the remaining bounding boxes, denoted as b_j, and eliminates any box that has an overlap with m exceeding a predefined Intersection over Union (IoU) threshold, T_n. This step is repeated iteratively, with the next highest-scoring box being selected and the overlap check being conducted until no boxes remain in the initial set [47].

$$score_j = \begin{cases} score_j, & if\ IoU\left(m, b_j\right) < T_n \\ 0, & if\ IoU\left(m, b_j\right) \geq T_n \end{cases} \quad (6)$$

Here, $score_j$ is the confidence score of the bounding box b_j. m is the bounding box with the highest confidence score that has been added to the final detection set. $IoU\left(m, b_j\right)$ measures the overlap between m and b_j. T_n is the IoU threshold used for suppressing overlapping boxes.

2) SOFT-NMS

Soft-NMS is an enhancement over traditional NMS that aims to refine object detection by addressing the limitations associated with the strict elimination of overlapping bounding boxes. Unlike NMS, which discards overlapping boxes

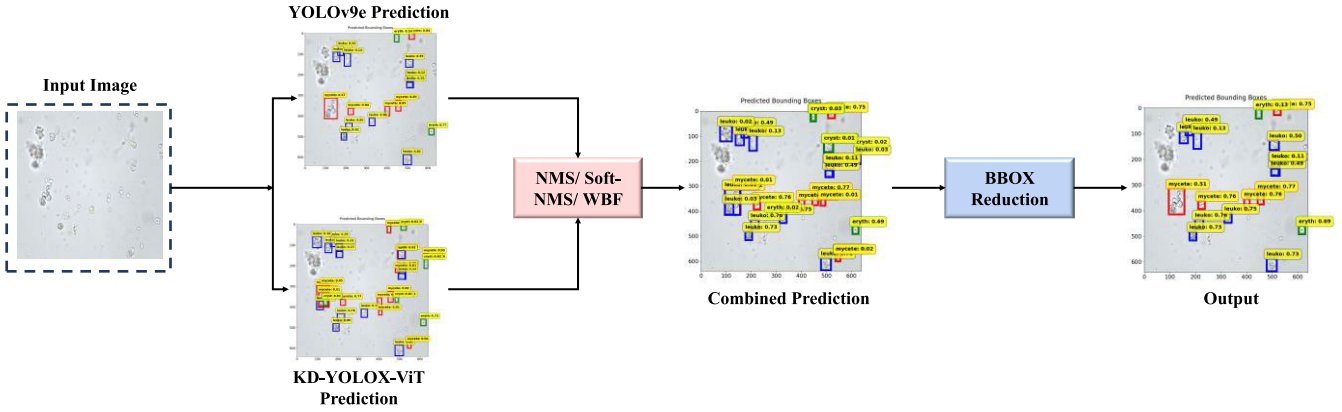

FIGURE 6. Block diagram illustrating the methodology of the ensembling method using YOLOv9-e and KD-YOLOX-ViT.

based on a fixed threshold, Soft-NMS adjusts the confidence scores of these boxes rather than completely removing them. This approach helps preserve nearby objects that may be close to each other, reducing the risk of false negatives. The Soft-NMS algorithm operates by applying a decaying function to the objectness scores of bounding boxes that overlap significantly with the highest-scoring box. Specifically, it reduces the score of each overlapping box in proportion to the degree of overlap, using a linear or Gaussian decay function. The effect of this adjustment is that boxes with minor overlaps maintain higher confidence scores, while those with substantial overlaps experience a more significant reduction in their scores [48].

$$score_j = \begin{cases} score_j \times \exp\left(-\dfrac{IoU(m, b_j)^2}{\sigma}\right), & if\ IoU\left(m, b_j\right) \geq T_s \\ score_j, & if\ IoU\left(m, b_j\right) < T_s \end{cases} \tag{7}$$

Here, $score_j$ is the adjusted confidence score of the bounding box b_j. m is the bounding box with the highest confidence score. $IoU\left(m, b_j\right)$ measures the overlap between m and b_j. T_s is the Soft-NMS threshold for the overlap. σ is a parameter that controls the extent of score reduction based on overlap.

3) WEIGHTED BOX FUSION (WBF)

To improve detection accuracy, Weighted Box Fusion (WBF) was adopted as an ensemble strategy alongside traditional post-processing techniques such as Non-Maximum Suppression (NMS) and Soft-NMS. In contrast to these suppression-based methods that eliminate redundant detections, WBF aggregates multiple bounding boxes that likely correspond to the same object by computing a confidence-weighted average of their coordinates. This fusion approach enables more comprehensive utilization of the information provided by individual detectors, particularly when predictions vary slightly in position and confidence.

Initially, predicted bounding boxes are grouped into clusters based on their Intersection over Union (IoU). Boxes with an IoU exceeding a threshold T_w (commonly set to 0.5) are considered part of the same cluster. For each cluster, a fused bounding box B_f is computed using a weighted average:

$$B_f = \sum_{i=1}^{n} w_i \,.\, B_i$$

where B_i represents the coordinates of the i^{th} bounding box in the cluster, and w_i is its associated weight. The weight w_i is determined based on the box's confidence score s_i and model reliability score m_i, and is defined as:

$$w_i = \frac{s_i \,.\, m_i}{\sum_{j=1}^{n} s_j \,.\, m_j}$$

This formulation ensures that bounding boxes with higher confidence and from more reliable models have greater influence on the final fused output. By retaining and integrating the spatial and score information of all contributing detections, WBF yields more precise and stable localization results, particularly in complex or cluttered scenes.

H. EXPERIMENTAL SETUP

The experiments were carried out in a cloud-based computing environment specifically designed for high-performance tasks. This setup was equipped with advanced GPUs that facilitated efficient processing and analysis of large datasets. The cloud infrastructure allowed for scalable and flexible resource allocation, which was crucial for training and evaluating our deep-learning model. Detailed technical specifications of this computing environment, including the hardware and software configurations, are provided in **Table 2**.

IV. RESULT AND ANALYSIS

A. INDIVIDUAL MODEL PERFORMANCE

Table 3 displays a performance analysis of different object detection models using metrics such as precision, recall,

TABLE 2. Experimental setup.

Component	Specification
CPU	13th Gen Intel® Core™ i7-13700KF @ 3.40 GHz
GPU	NVIDIA RTX 4090 (24 GB VRAM)
RAM	64 GB DDR5
Storage	2 TB SSD
Operating system	64-bit Microsoft Windows® 10
Software environment	Python 3.11, Torch 2.1.0, CUDA 11.8
Other Python libraries	Numpy, Pandas, Matplotlib, SciPy, OpenCV, Pillow, Albumentations, Torchvision, scikit-learn, Tensorboard

TABLE 3. Performance of different object detection models.

Model	Precision	Recall	mAP50	mAP50-95
GELAN-C	0.867	0.879	0.913	0.617
GELAN-E	0.879	0.881	0.914	0.619
RT-DETR-L	0.649	0.682	0.7	0.41
YOLOv5s	0.768	0.796	0.824	0.485
YOLOv5m	0.798	0.808	0.85	0.513
YOLOv5x	0.806	0.818	0.852	0.52
YOLOv8l	0.789	0.843	0.852	0.526
YOLOv8x	0.806	0.825	0.858	0.523
YOLOv9e	**0.885**	**0.881**	**0.922**	**0.623**
YOLOv10x	0.829	0.86	0.887	0.581
KD-YOLOX-ViT	0.86	0.88	0.867	0.533

mAP50 and mAP50-95. Out of all these models, RT-DETR-L exhibits the lowest performance, with a mAP50 of 0.70 and a mAP50-95 of 0.410. On the other hand, YOLOv9e demonstrates superior performance, attaining a mAP50 score of 0.922 and a mAP50-95 score of 0.623. The mAP50 of YOLOv9e is approximately 31.4% more than that of RT-DETR-L, while its mAP50-95 is 51.8% higher. The YOLOv10x model, which is the second highest model, achieves a mAP50 of 0.887 and a mAP50-95 of 0.581. Both mAP50 and mAP50-95 are 6.4% lower than the corresponding values achieved by the YOLOv9e model. The YOLOv8x model, ranked third in terms of performance, achieves a mAP50 of 0.858 and a mAP50-95 of 0.523. The mAP50 of YOLOv8x is 7.2% lower than that of YOLOv9e, and its mAP50-95 is 15.2% lower. However, it is worth noting that YOLOv9e has a precision of 0.885 and a recall of 0.881 which are also highest among all the models. KD-YOLOX-ViT achieves a mAP50 of 0.866, which is slightly higher than YOLOv8x's mAP50 of 0.858. **Figure 7** illustrates the qualitative result across various models.

In addition to the performance analysis, it is important to consider the computational complexity of the models to understand their practicality in real-world applications. **Table 4** presents the computational complexity of the various object detection models, including key metrics such as the number of parameters, training time per epoch, inference time per image, and GFLOPs (Giga Floating Point Operations).

TABLE 4. Computational complexity analysis.

Model	#Params (M)	Training Time (seconds/ Per Epoch)	Inference time (ms/Image)	GFLOPs
GELAN-C	25.3	81	14.0	103.2
GELAN-E	67.3	182	14.9	192.2
RT-DETR-L	32.8	128	5.3	108.0
YOLOv5s	9.1	16	1.3	23.8
YOLOv5m	25.1	55	3	64.4
YOLOv5x	97.2	128	7.8	246.9
YOLOv8l	43.6	81	4.9	165.4
YOLOv8x	68.2	130	7.4	258.2
YOLOv9e	69.4	197	20.2	244.9
YOLOv10x	31.7	128	6.6	171.1
KD-YOLOX-ViT	---	91	7.8	---

Among the models, YOLOv9e, with 69.4 million parameters, is relatively large, which contributes to its increased training time of 197 seconds per epoch. However, its performance gains, such as a mAP50 score of 0.922, justify this additional computational cost. In comparison, smaller models such as YOLOv5s, which has only 9.1 million parameters, exhibit faster training times (16 seconds per epoch) and lower inference times (1.3 ms/image), but with lower performance in terms of mAP50 (0.824) and mAP50-95 (0.485). Despite the relatively high computational cost, YOLOv9e maintains a competitive inference time of 20.2 ms/image, which is still efficient compared to other high-performing models. When considering GFLOPs, YOLOv9e requires 244.9 GFLOPs, reflecting its high computational demand. Nonetheless, the balance between computational complexity and detection accuracy makes YOLOv9e a suitable model for scenarios that prioritize both speed and performance.

Table 5 provides a detailed breakdown of the YOLOv9e model's performance across various categories of urine sediment particles, highlighting precision, recall, mAP50, and mAP50-95. Among the categories, erythrocytes (eryth) achieved the highest precision at 0.955, reflecting the model's strongest accuracy in detecting this class. In contrast, casts (cast) recorded the lowest precision at 0.793, indicating less reliable detection for this category. When it comes to recall, leukocytes (leuko) demonstrated the highest value at 0.943, showcasing the model's effectiveness in identifying this class across the majority of instances. On the other hand, erythrocytes (eryth) had the lowest recall at 0.898, though this is still relatively high. For mAP50, leukocytes (leuko) achieved the highest score of 0.969, signifying exceptional overall detection performance, while epithelial nucleus (epithn) had the lowest mAP50 at 0.908, highlighting some challenges in detecting this less frequent class. **Figure 8** demonstrates convergence curves of training process for YOLOv9 model.

B. EXTERNAL VALIDATION

External validation is an essential process for evaluating the robustness and generalizability of a deep learning model. By evaluating the model's performance on a separate dataset

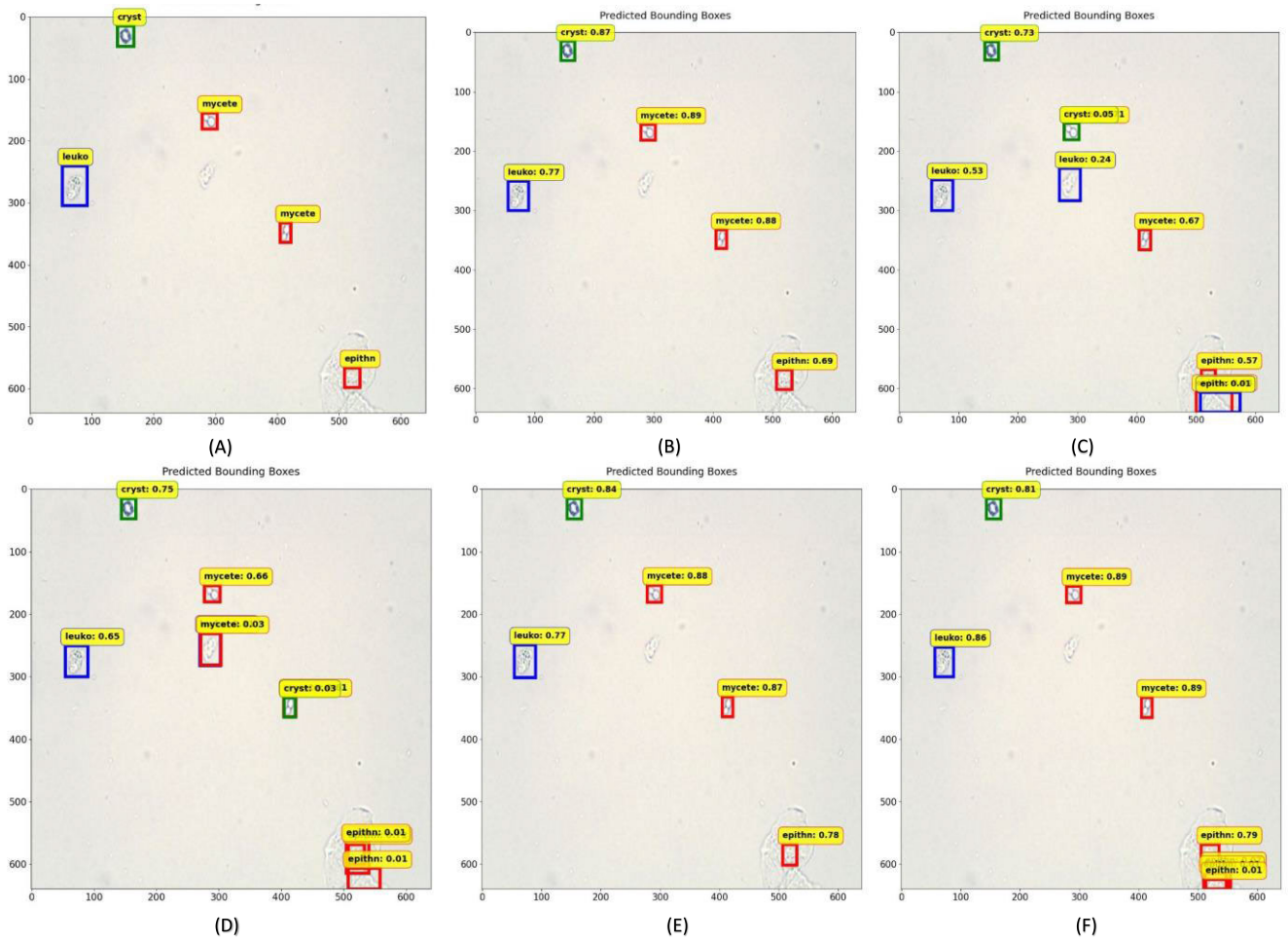

FIGURE 7. Individual model performance on sample test image. (A) Ground truth, (B) GELAN-E, (C) YOLOv5x, (D) YOLOv8x, (E) YOLOv9e,(F) YOLOv10x.

TABLE 5. Detailed analysis for each category of YOLOv9e.

Category	Instances	Precision	Recall	mAP50	mAP50-95
cast	545	0.793	0.771	0.843	0.568
cryst	317	0.894	0.875	0.933	0.645
epith	972	0.851	0.903	0.926	0.656
epithn	77	0.869	0.909	0.908	0.525
eryth	**3008**	**0.955**	0.898	0.943	0.632
leuko	796	0.912	**0.943**	**0.969**	**0.673**
mycete	233	0.923	0.87	0.934	0.663

that was not used for training, we can assess its ability to make accurate predictions on new and unknown data. This is crucial for assuring the reliability of the model across various datasets and clinical environments. This procedure aids in the identification of possible constraints or prejudices in the model, guaranteeing that it is not excessively tailored to the training data and can effectively apply to different datasets.

For external validation of the urine sediment detection model, a clinical microscopy dataset was selected. This dataset posed a new difficulty because the class labels differed from the labels of the training data obtained from the USE dataset. The training dataset included various classes, which are cast, cryst (crystals), epith (epithelial cells), epithn (epithelial nuclei), eryth (erythrocyte/RBC), leuko (leukocyte/WBC), and mycete. Conversely, the external validation dataset had categories such as Rod, RBC/WBC, Yeast, Miscellaneous, Single EPC, Small EPC sheet, and Large EPC sheet. In order to synchronize the external dataset with our model, we reassigned the labels of the external validation dataset. Specifically, we mapped RBC/WBC to eryth, and merged Single EPC, Small EPC sheet, and Large EPC sheet into epith. By reclassifying, we were able to utilize the trained YOLOv9e model on the external dataset. Nevertheless, the first findings indicated a decline in performance in **Table 6**, with an overall precision of 0.361, recall of 0.463, mAP50 of 0.375, and mAP50-95 of 0.21. The epith class demonstrated a precision of 0.102, recall of 0.144, and mAP50 of 0.0662. On the other hand, the eryth class exhibited superior performance, achieving a precision of 0.62, recall of 0.783, and mAP50 of 0.683. The lower mAP can be attributed to several factors. The epithelial cells in the USE training dataset exhibited distinct characteristics compared to those

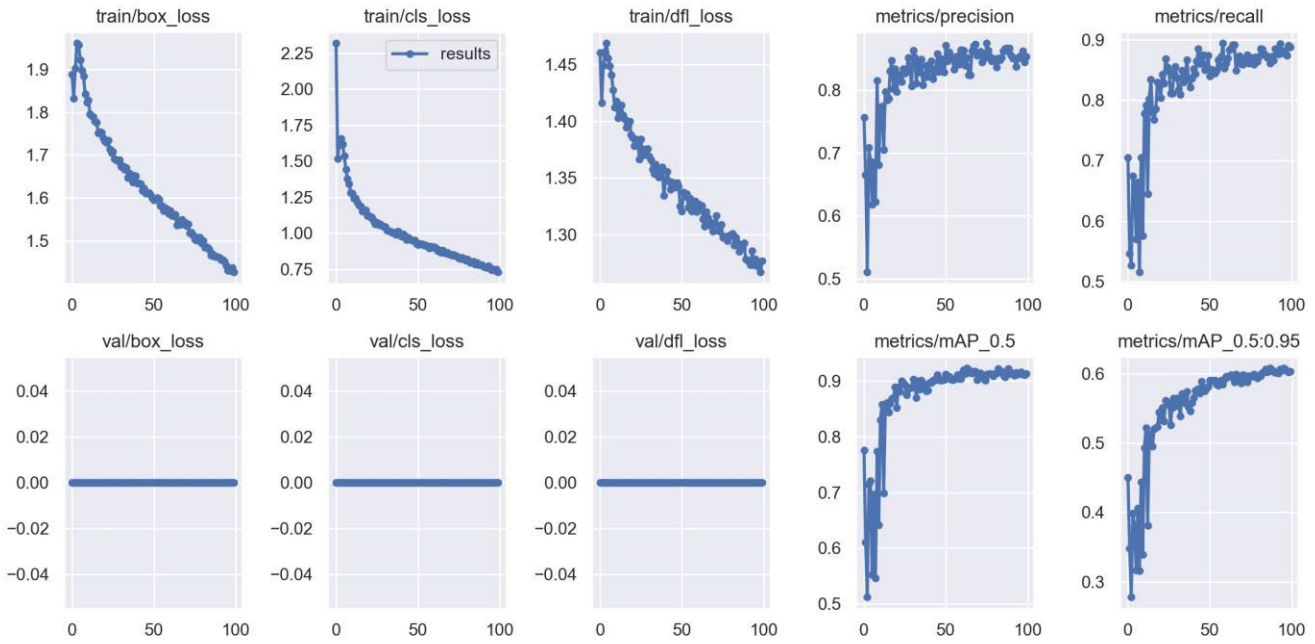

FIGURE 8. Convergence curves for the training process of the YOLOv9 model.

in the clinical microscopy dataset, resulting in challenges in accurately classifying them (**Figure 9(A)** and **9(B)**). Moreover, the clinical microscopy dataset merged red blood cells (RBC) and white blood cells (WBC) into a unified category (**Figure 9(C)** and **9(D)**), while the training dataset considered them as distinct categories, hence preventing precise detection. Moreover, several occurrences identified as epith in the external dataset were mistakenly categorized as epithn which is demonstrated in **Figure 9(E)** and **9(F)**, most likely because the training dataset included the epithn class.

To improve the model's generalizability, the clinical microscopy dataset was utilized for fine-tuning. The dataset was divided, with 20% allocated for training (60 images), 10% for validation (30 images), and 70% for testing (210 images). Following fine-tuning, a significant improvement in performance was observed. The test results post-tuning indicated an overall precision of 0.887, recall of 0.875, mAP50 of 0.899, and mAP50-95 of 0.675. The epith class achieved a precision of 0.87, recall of 0.913, mAP50 of 0.913, and mAP50-95 of 0.794, while the eryth class demonstrated a precision of 0.904, recall of 0.837, mAP50 of 0.885, and mAP50-95 of 0.556.

The second approach to external validation involved adopting a distinct strategy to enhance the alignment between the training USE dataset and the clinical microscopy dataset used for validation. **Table 7** demonstrates the result of second approach. The first phase consisted of reorganizing the training dataset by merging the class labels to align with those in the external dataset. To be more precise, the training set's classes epith (epithelial cells), eryth (erythrocytes/RBC), and leuko (leukocytes/WBC) were combined into two classes:

TABLE 6. Performance metrics for YOLOv9e model before and after fine-tuning on clinical microscopy dataset in first approach.

Initial Validation						
Class	**Images**	**Instances**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
epith	300	570	0.102	0.144	0.0662	0.0373
eryth (rbc/wbc)	300	2590	0.62	0.783	0.683	0.384
All	300	3160	0.361	0.463	0.375	0.21
After Fine-Tuning						
Class	**Images**	**Instances**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
epith	210	412	0.87	0.913	0.913	0.794
eryth (rbc/wbc)	210	2111	0.904	0.837	0.885	0.556
All	210	2523	**0.887**	**0.875**	**0.899**	**0.675**

*Bold value indicates the best results.

epith and rbc/wbc. The merging was also performed to align with the classes in the external dataset. In clinical microscopy dataset, RBC/WBC was associated with eryth, and the several epithelial cell classes (Single EPC, Small EPC sheet, Large EPC sheet) were combined into a single epith class. The external validation dataset then underwent preprocessing employing Contrast Limited Adaptive Histogram Equalization (CLAHE) and Gamma correction techniques to improve the quality of the images. The purpose of these preprocessing procedures was to normalize the images and enhance the

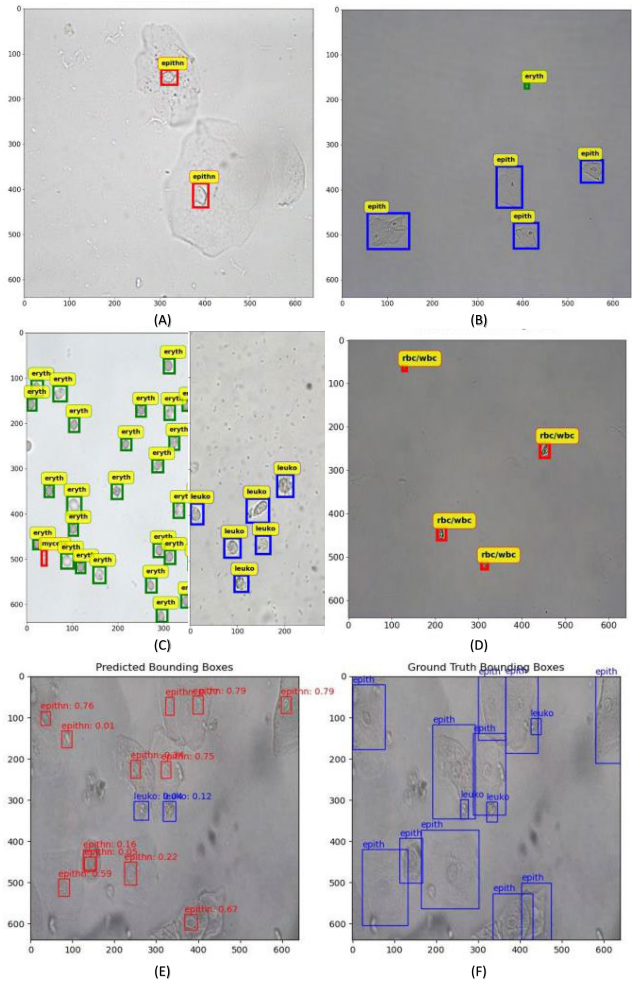

FIGURE 9. Difference between USE dataset and A Clinical Microscopy dataset. (A) Epithelial cell from USE dataset, (B) Epithelial cell from A Clinical Microscopy dataset, (C) RBC and WBC from USE dataset, (D) RBC/WBC from A Clinical Microscopy dataset, (E) Epithelial cell ground truth, (F) Model predicted epithelial cell as epithelial nuclei.

model's capacity to generalize across diverse datasets. However, only images that had enough, and similar occurrences were kept, resulting in a more polished dataset of 90 images.

When evaluated using this preprocessed dataset, the YOLOv9e model attained a mean Average Precision (mAP50) of 0.67, indicating a significant improvement of 78.67% compared to the initial approach (mAP50 = 0.375). Although there was an improvement, the overall performance was still below of the optimal level due to constant differences between the training and external validation datasets. In order to overcome these restrictions, the model underwent additional tuning. For this phase, the clinical microscopy dataset was partitioned into three subsets: 20% (60 images) for training, 10% (30 images) for validation, and 70% (210 images) for testing. In this instance, the fine-tuning procedure did not entail the removal of any images or the employing of CLAHE and Gamma correction. The findings following the fine-tuning process demonstrated a noteworthy enhancement in the performance of the model, achieving an overall precision of 0.849, recall of 0.902, mAP50 of 0.927, and mAP50-95 of 0.686. The epithelial class had a precision of 0.797, recall of 0.939, mean average precision at 50% overlap (mAP50) of 0.939, and mAP50-95 of 0.825. On the other hand, the red blood cell/white blood cell class attained a precision of 0.901, recall of 0.864, mAP50 of 0.915, and mAP50-95 of 0.548. The results were in line with previous findings, suggesting that the model had effectively adjusted to the external dataset without requiring any more preprocessing.

TABLE 7. Performance metrics for YOLOv9e model before and after fine-tuning on clinical microscopy dataset in second approach.

Initial Validation						
Class	**Images**	**Instances**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
epith	90	97	0.83	0.454	0.663	0.485
rbc/wbc	90	1997	0.74	0.718	0.678	0.38
All	90	2094	0.785	0.586	0.67	0.432
After Fine-Tuning						
Class	**Images**	**Instances**	**Precision**	**Recall**	**mAP50**	**mAP50-95**
epith	210	412	0.797	0.939	0.939	0.825
rbc/wbc	210	2111	0.901	0.864	0.915	0.548
All	210	2523	**0.849**	**0.902**	**0.927**	**0.686**

*Bold value indicates the best results.

The thorough external validation approach indicates that the YOLOv9e model for urine sediment analysis exhibits a substantial level of robustness and generalizability to apply to various environments. At first, the model's effectiveness was limited by differences in how classes were defined and the features of the images in the training and external validation datasets. However, by using systematic modifications such as reclassifying training data, implementing sophisticated preprocessing techniques, and fine-tuning the model, there was a significant enhancement in predicting accuracy for both epithelial and RBC/WBC categories. The model's capacity to adjust and improve its performance through these repetitive improvements highlights its potential for stable utilization in various clinical settings. The model's performance consistently aligns with the validation dataset, especially after fine-tuning without any further preprocessing. This clearly indicates that the YOLOv9e model is both robust and generalized for practical application in urine sediment identification.

It is important to note that the USE and clinical microscopy datasets are inherently different, with variations in class definitions and image features. Fine-tuning was necessary to adapt the model to the distinct characteristics of the external dataset, and this process led to significant performance enhancement. While this fine-tuning improves the model's ability to handle clinical data, it is crucial to clarify that such fine-tuning does not fully confirm the model's ability to generalize to completely unseen datasets that were not part of the training or fine-tuning process.

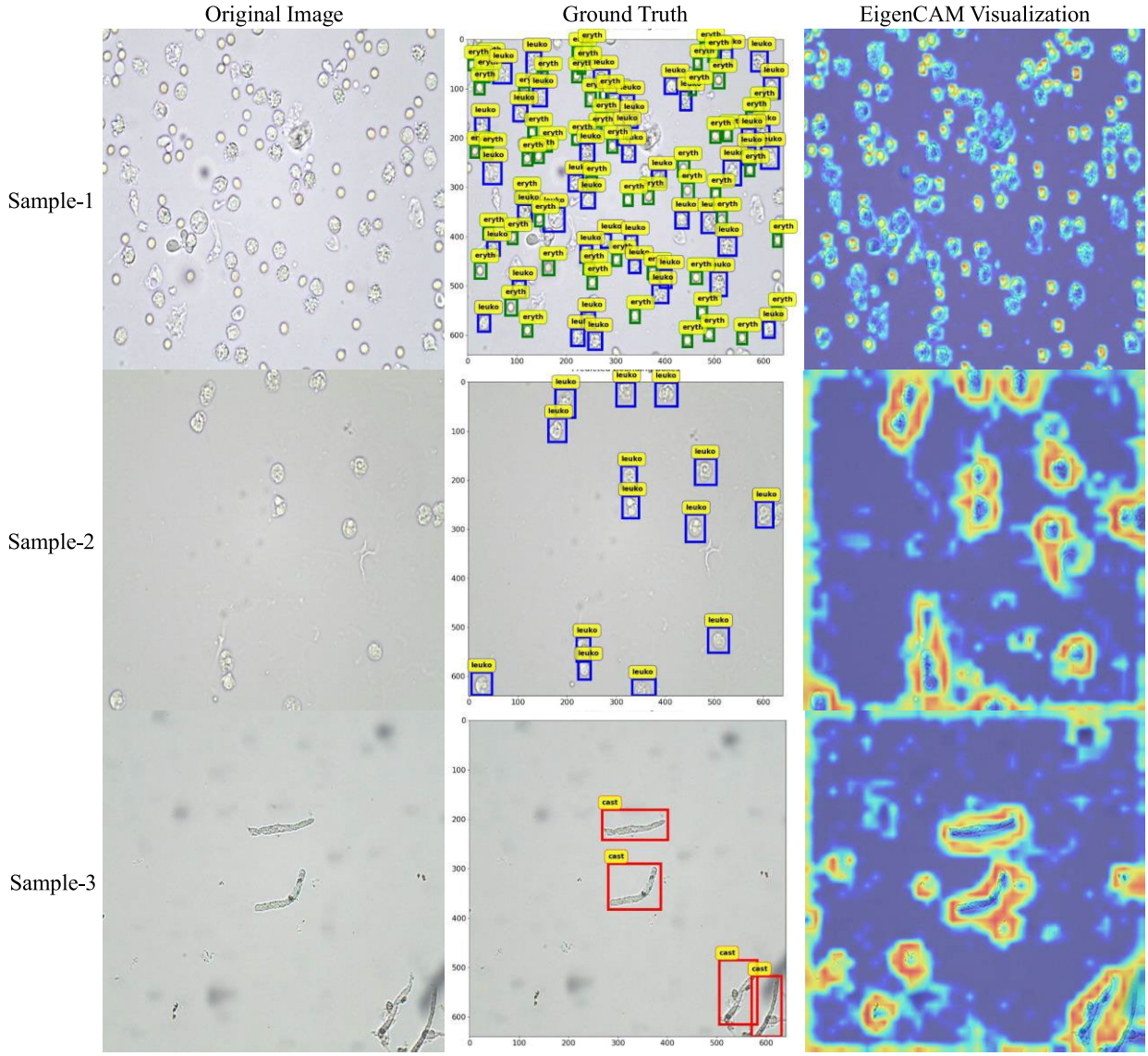

FIGURE 10. YOLOv9 EigenCAM visualization.

This validation highlights the model's adaptability and its promising potential for real-world applications, particularly when fine-tuned for specific clinical environments. However, to truly establish its generalization capability, further validation across additional, diverse, and previously unseen similar datasets is necessary. These future studies will help comprehensively assess the model's ability to function effectively in real-world clinical practice beyond the datasets used for training and fine-tuning.

C. INTERPRETABILITY

EigenCAM was utilized to improve the clarity of the deep learning model's predictions. The presented sample images in **Figure 10** illustrate the model's particular focus on important areas that are associated with sediments particles. The EigenCAM outputs highlight the locations that have a substantial impact on the model's detection process, while less important parts are shown by cooler shades. This interpretability approach validates the efficiency of the model in identifying significant features in the sediment samples, while minimizing the influence of irrelevant areas. This further strengthens the dependability of the detection outcomes.

D. RESULT OF ENSEMBLE METHOD

To determine their effect on object detection quality, the performance of ensemble approaches integrating initial predictions of YOLOv8x, KD-YOLOX-ViT, and YOLOv9e models was assessed. **Table 8** and **Figure 11** demonstrate the overall performance of each model and their ensemble results. Among the standalone models, YOLOv8x achieved a mAP50 of 85.8%, KD-YOLOX-ViT obtained a slightly higher score of 86.7%, and YOLOv9e significantly outperformed both with a mAP50 of 92.2%, indicating its superior object detection capability.

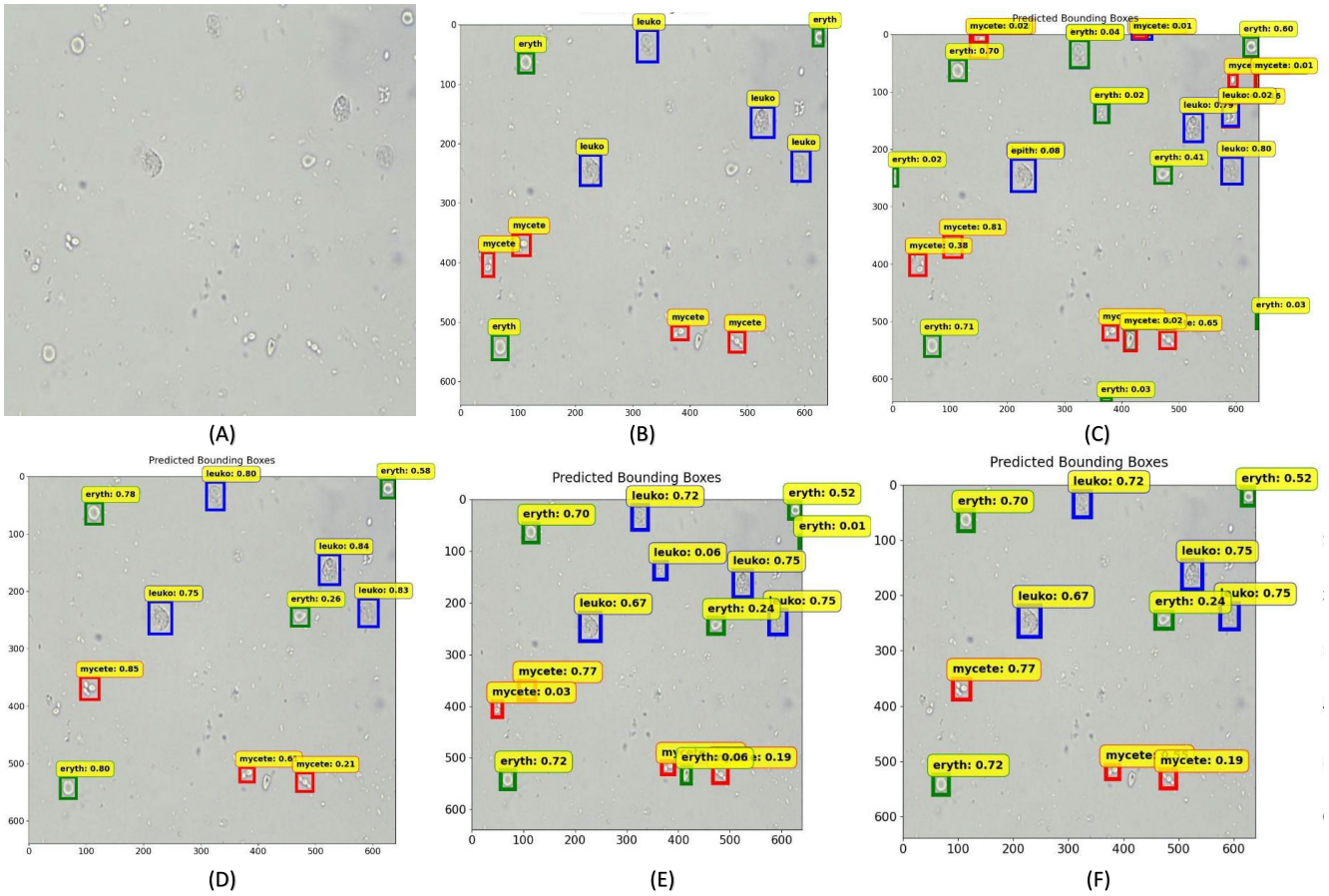

FIGURE 11. Predictions on sample test images using different ensemble methods: (A) Input image, (B) Ground truth, (C) KD-YOLOX-ViT,(D) YOLOv9e, (E) NMS, (F) WBF.

TABLE 8. Performance comparison of individual models and ensemble methods.

Model	Ensemble Method	mAP50 (%)
YOLOv8x	---	85.8
KD-YOLOX-ViT	---	86.7
YOLOv9e	---	92.2
YOLOv8x + KD-YOLOX-ViT	Soft-NMS	87.93
	NMS	88.33
	WBF	91.1
YOLOv8x + YOLOv9e	Soft-NMS	92.33
	NMS	92.63
	WBF	93.01
KD-YOLOX-ViT + YOLOv9e	Soft-NMS	93.51
	NMS	93.7
	WBF	**94.18**

*Bold value indicates the best results.

When YOLOv8x was combined with KD-YOLOX-ViT, a modest improvement was observed across all ensemble techniques. Soft-NMS yielded a mAP50 of 87.93%, while traditional NMS improved the performance to 88.33%. The use of Weighted Box Fusion (WBF), which considers both box coordinates and confidence scores for merging predictions, further elevated the performance to 91.1%. This demonstrates the benefit of combining complementary models even when their individual performance is lower. The ensemble of YOLOv8x and YOLOv9e produced even more notable results, with Soft-NMS achieving a mAP50 of 92.33% and NMS slightly surpassing that with 92.63%. WBF provided the best performance in this group at 93.01%, confirming its effectiveness in refining overlapping predictions. The highest detection performance was observed when KD-YOLOX-ViT and YOLOv9e were combined. In this setup, Soft-NMS achieved a mAP50 of 93.51%, NMS reached 93.7%, and WBF produced the best result overall with a mAP50 of 94.18%. This combination benefits from the strengths of both models—KD-YOLOX-ViT's transformer-based architecture and YOLOv9e's advanced backbone—which together contribute to more robust and accurate object detection.

Overall, the findings underscore the value of ensemble learning in object detection tasks. Soft-NMS offers incremental improvements by refining overlapping predictions, while NMS provides stronger filtering through confident box suppression. WBF consistently delivered the highest performance across all model combinations by effectively merging detections from multiple sources. Notably, the combination of KD-YOLOX-ViT and YOLOv9e using WBF achieved the highest mAP50 score of 94.18%, outperforming all

other ensembles. This demonstrates the strength of integrating transformer-based architectures with advanced YOLO backbones, further enhanced by the robust fusion capabilities of WBF. These results highlight the effectiveness of ensemble techniques in leveraging the diversity of individual model predictions to boost detection accuracy and robustness. The superior performance of the KD-YOLOX-ViT + YOLOv9e + WBF ensemble supports its recommendation as the optimal strategy for improving object detection performance in this study.

E. EXTARNAL VALIDATION USING ENSEMBLE METHOD (YOLOV9E + KD-YOLOX-VIT + WBF)

To ensure consistency and fair comparison, the ensemble model was evaluated using the same external validation strategy as previously applied for YOLOv9e. Specifically, label harmonization was performed by merging epith, eryth, and leuko classes into two categories: epith and rbc/wbc, aligning with the class definitions in the clinical microscopy dataset. The dataset used for final testing consisted of 210 clinical microscopy images without any additional preprocessing.

TABLE 9. External validation performance of ensemble method (YOLOv9e + KD-YOLOX-ViT + WBF).

Class	Images	Instances	mAP50
epith	210	412	0.9441
rbc/wbc	210	2111	0.9487
All	210	2523	**0.9464**

Table 9 presents the external validation result. The ensemble model (YOLOv9e + KD-YOLOX-ViT with Weighted Box Fusion) demonstrated superior performance. It achieved a final mAP50 of 0.9464, showing improvement over the individual models. Class-wise, the epith category attained an average precision (AP) of 0.9441 with 412 ground truth instances, while the rbc/wbc category reached an AP of 0.9487 across 2111 instances. These results highlight the ensemble model's strong generalization capability and robustness across heterogeneous clinical microscopy data.

F. ABLATION STUDY

The ablation study was performed in order to rigorously assess the contribution of different components and methodologies to the overall performance of the urine sediment detection model, offering insights into the influence of each element on the accuracy of detection. This study is essential for evaluating the efficacy of particular model configurations and optimization methodologies.

1) IMPACT OF IMAGE ENHANCEMENT TECHNIQUE ON MODEL PERFORMANCE

Table 10 provides information on a study that investigates the impact of several image enhancement techniques on the performance metrics of the YOLOv9e model. The evaluated strategies include Histogram Equalization, CLAHE (Contrast Limited Adaptive Histogram Equalization) combined with Gamma Correction, and a baseline scenario without any enhancement. Applying Histogram Equalization yielded an accuracy of 85.9% and a sensitivity of 87.0%, accompanied by a mAP50 of 90.4% and a mAP50-95 of 59.1%. This method, which enhances the difference in brightness levels of an image by dispersing them, has significantly improved the accuracy of detecting objects, especially in regions where their characteristics were previously less clear. Nevertheless, the comparatively lower mAP50-95 indicates that although the model exhibited great performance at higher IoU thresholds, it demonstrated poorer consistency across a wider range of thresholds. On the other hand, when CLAHE and Gamma Correction were used together, there were additional enhancements, resulting in a precision of 86.8%, a recall of 88.4%, and a mAP50 of 92.1%. The mAP50-95 also improved to 60.0%. The combination of CLAHE, a technique that improves image contrast by applying local histogram equalization, and Gamma Correction, which adjusts image brightness, resulted in a more balanced improvement. This methodology enabled the model to continuously achieve high performance across various detection thresholds.

TABLE 10. Performance comparison of YOLOv9e with different image enhancement techniques.

Image Enhancement Technique	Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
Histogram Equalization	YOLOv9e	85.9	87.0	90.4	59.1
CLAHE and Gamma Correction		86.8	**88.4**	92.1	60.0
----		**88.5**	88.1	**92.2**	**62.3**

*Bold value indicates the best results.

However, the baseline model, without any improvements, achieved the highest performance measures. It had a precision of 88.5%, a recall of 88.1%, a mAP50 of 92.2%, and a mAP50-95 of 62.3%. This result indicates that although improvement strategies can enhance some parts of model performance, the default image quality and contrast may already be optimal for YOLOv9e's detection capabilities. The better baseline scores indicate that the model's structure and training data are well-matched for the detection job, reducing the necessity for extra improvement strategies in this particular situation.

2) ENSEMBLE OF DIFFERENT MODELS

This study evaluated various combinations of top performing models—YOLOv8x, YOLOv9e, YOLOv10x, and KD-YOLOX-ViT—using ensemble methods such as Soft-NMS, NMS, and WBF. As shown in **Table 11**, the results indicate that WBF consistently outperformed the other two ensemble strategies across most model combinations. For example, the YOLOv10x and YOLOv9e ensemble achieved 93.3% with WBF, surpassing the NMS and Soft-NMS variants. Similarly, the YOLOv8x and KD-YOLOX-ViT

combination reached 91.1% using WBF, compared to 87.93% with Soft-NMS and 88.33% with NMS.

Among all combinations, the ensemble of KD-YOLOX-ViT and YOLOv9e using WBF achieved the highest mAP50 of 94.18%, establishing it as the best-performing setup in this study. According to these findings, WBF emerges as a superior ensemble technique due to its ability to merge bounding boxes more effectively rather than suppressing them. Moreover, the complementary strengths of KD-YOLOX-ViT and YOLOv9e in capturing fine-grained features and robust spatial representations make this pairing particularly powerful. Based on this comprehensive comparison, the KD-YOLOX-ViT + YOLOv9e ensemble with WBF is selected as the final model for deployment.

TABLE 11. Performance comparison of different YOLO model combinations using ensemble methods.

Model Combination	Ensemble Method	mAP50 (%)
YOLOv8x + YOLOv9e	Soft-NMS	92.33
	NMS	92.63
	WBF	93.01
YOLOv8x + YOLOv10x	Soft-NMS	90.24
	NMS	90.86
	WBF	88.25
YOLOv8x + KD-YOLOX-ViT	Soft-NMS	87.93
	NMS	88.33
	WBF	91.1
YOLOv10x + YOLOv9e	Soft-NMS	92.3
	NMS	92.53
	WBF	93.3
KD-YOLOX-ViT + YOLOv9e	Soft-NMS	93.51
	NMS	93.7
	WBF	**94.18**
YOLOv10x + KD-YOLOX-ViT	Soft-NMS	87.9
	NMS	88.33
	WBF	89.30

*Bold value indicates the best results.

G. DISCUSSION, LIMITATIONS AND FUTURE WORK

The YOLOv9e and KD-YOLOX-ViT models were employed for urine sediment detection, and a thorough external validation process was carried out utilizing an entirely new data set to assess the model's adaptability for real-world situations. This external validation process demonstrates the model's potential for real-world application, though it is important to note that the validation was conducted on a specific dataset that differs from those used during training. While this indicates the model's adaptability to some variations in data, further validation across a broader range of previously unseen, similar real-world datasets is necessary to fully assess its generalization capabilities and efficacy in diverse clinical contexts. In order to improve the accuracy, the YOLOv9e model was combined with KD-YOLOX-ViT using an ensemble technique. Weighted Box Fusion (WBF) was utilized to refine the detection outcomes by merging overlapping bounding boxes from different models based on their confidence scores, rather than suppressing them. This method preserves more useful information, leading to enhanced accuracy, reduced false positives, and over-all more reliable detection results. The utilization of EigenCAM yielded substantial benefits in regards to the interpretability of the model. EigenCAM provided coherent insights into the decision-making process by showing the specific regions of the input images that influenced the model's predictions. This transparency allows for better understanding of how the model interprets data, which can facilitate trust in the model's predictions and support diagnostic decisions in clinical settings.

Notable performance metrics were achieved with the YOLOv9e model in this study. YOLOv9e had a mAP50 of 92.2%, indicating an impressive amount of accuracy in identifying different sediment classes. The combination of YOLOv9e and KD-YOLOX-ViT using WBF ensemble technique led to a rise in the mAP50 to 94.18%. This improvement highlights the efficacy of the ensemble method, which utilizes the advantages of both models to attain higher performance. The performance of the ensemble technique was evaluated and compared with several state-of-the-art (SOTA) models as indicated in**Table 12**. The findings reveal that this method outperformed nearly all of the current models, with the exception of the YOLOX-based model (YUS-Net) [12], which attained a higher mAP50 of 96.07%. While the YUS-Net model, based on YOLOX, reports impressive mAP50 values exceeding 99% for certain sediment classes such as ''cast'', ''leukocyte'', and ''mycete'', and even 100% for ''cryst'' and ''epithn'', these results seem unusually high and may not fully represent the challenges faced in real-world clinical microscopic sediment detection. Such near-perfect performance is rarely observed in clinical microscopic sediment detection tasks, where variability in image quality, lighting conditions, and staining techniques often complicate model accuracy. Furthermore, the YUS-Net study does not provide sufficient evidence of the model's generalization to diverse, unseen clinical datasets. In contrast, our study presents a more comprehensive evaluation that includes both the USE dataset and an external clinical microscopy dataset. While our reported mAP50 values (92.2% for YOLOv9e and 92.63% for the ensemble model) are slightly lower, they reflect a more realistic and generalized performance, especially when considering the model's adaptability to different clinical settings. The absence of external validation and the exceedingly high results in the YUS-Net study raise concerns about the model's robustness and its practical applicability across a range of real-world clinical environments.

Although the proposed approach did not achieve the maximum mAP50, it still offers substantial improvements in terms of its capacity to generalize and interpret results. Unlike other SOTA models, which may lack extensive validation and interpretability features, the YOLOv9e model and YOLOv9e + KD-YOLOX-ViT + WBF was rigorously validated with an external dataset and provided interpretability through EigenCAM. This comprehensive approach ensures that the proposed model's predictions are not only accurate but also transparent and reliable, making the

TABLE 12. Performance comparison with the state-of-the-art methods.

Year	Ref.	Dataset	Classes	Method	Performance
2018	[11]	USE (5377 images)	7 classes	DenseNet with a Feature Pyramid Network (FPN) DFPN	mAP: 86.9%
2018	[13]	USE (5377 images)	7 classes	Faster R-CNN and Single Shot MultiBox Detector (SSD)	mAP: 84.1%
2020	[15]	Private Dataset (15360 images)	7 classes	Feature extractor: ResNet50 Detection: FPN	mAP: 88.6%
2023	[12], [17]	USE (5377 images)	7 classes	YOLOX-based model (YUS-Net)	mAP: 96.07%
2023	[19]	USE (5377 images)	7 classes	YOLOv5 with Evolutionary Genetic Algorithm (EGA)	YOLOv5l mAP: 85.8% YOLOv5x mAP: 85.4%
	Proposed	USE (5377 images), A clinical microscopy dataset (300 images)	7 classes	YOLOv9e + KD-YOLOX-ViT + WBF	mAP: 94.18%

proposed method robust and generalized for urine sediment detection.

This study admits several constraints and limitations. The biggest constraint is the slightly reduced accuracy in comparison to the top-performing model, however the accuracy still stays competitive within the field. Furthermore, the size of the external validation dataset employed in this study is quite limited, which could affect the strength of the assertions regarding generalizability. The use of the KD-YOLOX-ViT model in the ensemble may contribute to performance variability, as it can affect the results based on the specific properties of the datasets used. Future research should focus on several key areas to build upon the findings of this study. Increasing the size of the external validation dataset to include a larger variety of samples would yield a more thorough evaluation of the generalizability of the model. A wider range of models and the investigation of sophisticated ensemble techniques may be included to improve performance even more. Urine sediment detection systems could be made more transparent and reliable by looking into alternate interpretability strategies and combining them with other cutting-edge models. Furthermore, the utilization of the established techniques on different clinical datasets and scenarios will provide an evaluation of the flexibility and efficacy of the suggested method in diverse contexts.

V. CONCLUSION

This study employed an advanced ensemble deep learning framework, precisely the YOLOv9e + KD-YOLOX-ViT + WBF model, to automate the identification and categorization of urine sediment particles from microscopic images. Our approach aims to meet the acute need for prompt and precise urinalysis, especially in situations where there is limited availability of professional urology experts and equipment. The ensemble model exhibited outstanding performance in detecting various categories, indicating its ability to greatly improve diagnostic accuracy and efficiency. Our research shows that the model surpasses numerous cutting-edge techniques, providing faster detection and higher classification accuracy. The utilization of ensemble DL method in urine sediment research has great potential in optimizing clinical workflows, facilitating early disease identification, and ultimately enhancing therapeutic outcomes. The adoption of such technologies in pediatric urology and broader urinalysis could greatly reduce the workload of healthcare professionals, expedite diagnostic processes, and ensure timely medical intervention. Continued refinement and application of these techniques are essential for advancing diagnostic practices and expanding access to high-quality care.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The data used in this study is available upon reasonable request to the corresponding author.

CONFLICT OF INTEREST

The authors declare no conflicts of interest for this study.

REFERENCES

[1] X. Wu, Y. Dong, Y. Liu, Y. Li, Y. Sun, J. Wang, and S. Wang, "The prevalence and predictive factors of urinary tract infection in patients undergoing renal transplantation: A meta-analysis," *Amer. J. Infection Control*, vol. 44, no. 11, pp. 1261–1268, Nov. 2016, doi: 10.1016/j.ajic.2016.04.222.

[2] H. C. Arora, M. Fascelli, J. H. Zhang, S. Isharwal, and S. C. Campbell, "Kidney, ureteral, and bladder cancer," *Med. Clinics North Amer.*, vol. 102, no. 2, pp. 231–249, Mar. 2018, doi: 10.1016/j.mcna.2017.10.002.

[3] X. Yang, H. Chen, Y. Zheng, S. Qu, H. Wang, and F. Yi, "Disease burden and long-term trends of urinary tract infections: A worldwide report," *Frontiers Public Health*, vol. 10, Jul. 2022, Art. no. 888205, doi: 10.3389/fpubh.2022.888205.

[4] L. Zhang, L. Zhu, T. Xu, J. Lang, Z. Li, J. Gong, Q. Liu, and X. Liu, "A population-based survey of the prevalence, potential risk factors, and symptom-specific bother of lower urinary tract symptoms in adult Chinese women," *Eur. Urol.*, vol. 68, no. 1, pp. 97–112, Jul. 2015, doi: 10.1016/j.eururo.2014.12.012.

[5] P. Bafna, S. Deepanjali, J. Mandal, N. Balamurugan, R. P. Swaminathan, and T. Kadhiravan, "Reevaluating the true diagnostic accuracy of dipstick tests to diagnose urinary tract infection using Bayesian latent class analysis," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0244870, doi: 10.1371/journal.pone.0244870.

[6] R. Barata, D. Navarro, N. Moreira Fonseca, A. Carina Ferreira, M. Góis, H. Viana, and F. Nolasco, "MO049: Correlation of findings in urinary sediment microscopy and histological lesions in kidney biopsy: Red flags not to be missed," *Nephrol. Dialysis Transplantation*, vol. 37, pp. 30–48, May 2022, doi: 10.1093/ndt/gfac063.001.

[7] C. Cavanaugh and M. A. Perazella, "Urine sediment examination in the diagnosis and management of kidney disease: Core curriculum 2019," *Amer. J. Kidney Diseases*, vol. 73, no. 2, pp. 258–272, Feb. 2019, doi: 10.1053/j.ajkd.2018.07.012.

[8] G. J. Kost, N. K. Tran, and R. F. Louie, "Point-of-Care testing: Principles, practice, and critical-emergency-disaster medicine," in *Encyclopedia of Analytical Chemistry*. Hoboken, NJ, USA: Wiley, 2008, doi: 10.1002/9780470027318.a0540.pub2.

[9] H. A. Boon, T. De Burghgraeve, J. Y. Verbakel, and A. Van den Bruel, "Point-of-care tests for pediatric urinary tract infections in general practice: A diagnostic accuracy study," *Family Pract.*, vol. 39, no. 4, pp. 616–622, Jul. 2022, doi: 10.1093/fampra/cmab118.

[10] G. J. Becker, G. Garigali, and G. B. Fogazzi, "Advances in urine microscopy," *Amer. J. Kidney Diseases*, vol. 67, no. 6, pp. 954–964, Jun. 2016, doi: 10.1053/j.ajkd.2015.11.011.

[11] Y. Liang, Z. Tang, M. Yan, and J. Liu, "Object detection based on deep learning for urine sediment examination," *Biocybern. Biomed. Eng.*, vol. 38, no. 3, pp. 661–670, 2018, doi: 10.1016/j.bbe.2018.05.004.

[12] M. Yu, Y. Lei, W. Shi, Y. Xu, and S. Chan, "An improved YOLOX for detection in urine sediment images," in *Proc. Int. Conf. Intell. Robot. Appl.*, 2022, pp. 556–567, doi: 10.1007/978-3-031-13841-6_50.

[13] Y. Liang, R. Kang, C. Lian, and Y. Mao, "An end-to-end system for automatic urinary particle recognition with convolutional neural network," *J. Med. Syst.*, vol. 42, no. 9, p. 165, Sep. 2018, doi: 10.1007/s10916-018-1014-6.

[14] S. Chan, B. Wu, H. Wang, X. Zhou, G. Zhang, and G. Wang, "Cross-domain mechanism for few-shot object detection on urine sediment image," *Comput. Biol. Med.*, vol. 166, Nov. 2023, Art. no. 107487, doi: 10.1016/j.compbiomed.2023.107487.

[15] Q. Li, Z. Yu, T. Qi, L. Zheng, S. Qi, Z. He, S. Li, and H. Guan, "Inspection of visible components in urine based on deep learning," *Med. Phys.*, vol. 47, no. 7, pp. 2937–2949, Jul. 2020, doi: 10.1002/mp.14118.

[16] D. Avci, E. Sert, E. Dogantekin, O. Yildirim, R. Tadeusiewicz, and P. Plawiak, "A new super resolution faster R-CNN model based detection and classification of urine sediments," *Biocybern. Biomed. Eng.*, vol. 43, no. 1, pp. 58–68, Jan. 2023, doi: 10.1016/j.bbe.2022.12.001.

[17] H. Lyu, F. Xu, T. Jin, S. Zheng, C. Zhou, Y. Cao, B. Luo, Q. Huang, W. Xiang, and D. Li, "Automated detection of multi-class urinary sediment particles: An accurate deep learning approach," *Biocybern. Biomed. Eng.*, vol. 43, no. 4, pp. 672–683, Oct. 2023, doi: 10.1016/j.bbe.2023.09.003.

[18] Z. Chen, R. Hu, F. Chen, H. Fan, F. Y. Ching, Z. Li, and S. Su, "An efficient particle YOLO detector for urine sediment detection," in *Proc. Int. Conf. Mach. Learn. Cyber Secur.*, 2023, pp. 294–308, doi: 10.1007/978-3-031-20102-8_23.

[19] K. Suhail and D. Brindha, "Microscopic urinary particle detection by different YOLOv5 models with evolutionary genetic algorithm based hyperparameter optimization," *Comput. Biol. Med.*, vol. 169, Feb. 2024, Art. no. 107895, doi: 10.1016/j.compbiomed.2023.107895.

[20] Q. Ji, X. Li, Z. Qu, and C. Dai, "Research on urine sediment images recognition based on deep learning," *IEEE Access*, vol. 7, pp. 166711–166720, 2019, doi: 10.1109/ACCESS.2019.2953775.

[21] K. Suhail and D. Brindha, "A review on various methods for recognition of urine particles using digital microscopic images of urine sediments," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102806, doi: 10.1016/j.bspc.2021.102806.

[22] M. Erten, P. D. Barua, I. Tuncer, S. Dogan, M. Baygin, T. Tuncer, R.-S. Tan, and U. R. Acharya, "Swin-LBP: A competitive feature engineering model for urine sediment classification," *Neural Comput. Appl.*, vol. 35, no. 29, pp. 21621–21632, Oct. 2023, doi: 10.1007/s00521-023-08919-w.

[23] X. Zhang, L. Jiang, D. Yang, J. Yan, and X. Lu, "Urine sediment recognition method based on multi-view deep residual learning in microscopic image," *J. Med. Syst.*, vol. 43, no. 11, p. 325, Nov. 2019, doi: 10.1007/s10916-019-1457-4.

[24] M. Yildirim, H. Bingol, E. Cengil, S. Aslan, and M. Baykara, "Automatic classification of particles in the urine sediment test with the developed artificial intelligence-based hybrid model," *Diagnostics*, vol. 13, no. 7, p. 1299, Mar. 2023, doi: 10.3390/diagnostics13071299.

[25] M. Yan, Q. Liu, Z. Yin, D. Wang, and Y. Liang, "A bidirectional context propagation network for urine sediment particle detection in microscopic images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 981–985, doi: 10.1109/ICASSP40776.2020.9054367.

[26] N. Liou, T. De, A. Urbanski, C. Chieng, Q. Kong, A. L. David, R. Khasriya, A. Yakimovich, and H. Horsley, "A clinical microscopy dataset to develop a deep learning diagnostic test for urinary tract infection," *Sci. Data*, vol. 11, no. 1, p. 155, Feb. 2024, doi: 10.1038/s41597-024-02975-0.

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.

[28] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, and J.-S. Chiang, "YOLOv9 for fracture detection in pediatric wrist trauma X-ray images," *Electron. Lett.*, vol. 60, no. 11, pp. 1–11, Jun. 2024, doi: 10.1049/ell2.13248.

[29] Y. Li, Z. Hu, Y. Zhang, J. Liu, W. Tu, and H. Yu, "DDEYOLOv9: Network for detecting and counting abnormal fish behaviors in complex water environments," *Fishes*, vol. 9, no. 6, p. 242, Jun. 2024, doi: 10.3390/fishes9060242.

[30] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A new backbone that can enhance learning capability of CNN," 2019, *arXiv:1911.11929*.

[31] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," 2022, *arXiv:2211.04800*.

[32] S. Kittisorayut, S. Nonsiri, P. Kamin, and S. Makdee, "A machine learning algorithms for drug identification: Enhancing medication safety in Thailand," in *Proc. 21st Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jun. 2024, pp. 47–52, doi: 10.1109/jcsse61278.2024.10613653.

[33] Ö. Kaya, M. Y. Çodur, and E. Mustafaraj, "Automatic detection of pedestrian crosswalk with faster R-CNN and YOLOv7," *Buildings*, vol. 13, no. 4, p. 1070, Apr. 2023, doi: 10.3390/buildings13041070.

[34] N. Chandra, H. Vaidya, S. Sawant, and S. R. Meena, "A novel attention-based generalized efficient layer aggregation network for landslide detection from satellite data in the higher himalayas, Nepal," *Remote Sens.*, vol. 16, no. 14, p. 2598, Jul. 2024, doi: 10.3390/rs16142598.

[35] M. G. Ragab, S. J. Abdulkadir, A. Muneer, A. Alqushaibi, E. H. Sumiea, R. Qureshi, S. M. Al-Selwi, and H. Alhussian, "A comprehensive systematic review of YOLO for medical object detection (2018 to 2023)," *IEEE Access*, vol. 12, pp. 57815–57836, 2024, doi: 10.1109/access.2024.3386826.

[36] M. Aubard, L. Antal, A. Madureira, and E. Àbrahám, "Knowledge distillation in YOLOX-ViT for side-scan sonar object detection," 2024, *arXiv:2403.09313*.

[37] M. Aubard, A. Madureira, L. Madureira, and J. Pinto, "Real-time automatic wall detection and localization based on side scan sonar images," in *Proc. IEEE/OES Auto. Underwater Vehicles Symp. (AUV)*, Sep. 2022, pp. 1–6, doi: 10.1109/AUV53081.2022.9965813.

[38] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.

[40] Y. Li, H. Wang, L. M. Dang, T. N. Nguyen, D. Han, A. Lee, I. Jang, and H. Moon, "A deep learning-based hybrid framework for object detection and recognition in autonomous driving," *IEEE Access*, vol. 8, pp. 194228–194239, 2020, doi: 10.1109/ACCESS.2020.3033289.

[41] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205, doi: 10.1109/ICCV.2019.00929.

[42] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[43] T. Rahman, A. Khandakar, K. R. Islam, M. M. Soliman, M. T. Islam, A. Elsayed, Y. Qiblawey, S. Mahmud, A. Rahman, F. Musharavati, E. Zalnezhad, and M. E. H. Chowdhury, "HipXNet: Deep learning approaches to detect aseptic loos-ening of hip implants using X-ray images," *IEEE Access*, vol. 10, pp. 53359–53373, 2022, doi: 10.1109/ACCESS.2022.3173424.

[44] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 2023, *arXiv:2304.00501*.

[45] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," 2020, *arXiv:2008.00299*.
[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[47] R. Sarmun, M. E. H. Chowdhury, M. Murugappan, A. Aqel, M. Ezzuddin, S. M. Rahman, A. Khandakar, S. Akter, R. Alfkey, and A. Hasan, "Diabetic foot ulcer detection: Combining deep learning models for improved localization," *Cognit. Comput.*, vol. 16, no. 3, pp. 1413–1431, May 2024, doi: 10.1007/s12559-024-10267-3.
[48] F. Bushra, M. E. H. Chowdhury, R. Sarmun, S. Kabir, M. Said, S. B. Zoghoul, A. Mushtak, I. Al-Hashimi, A. Alqahtani, and A. Hasan, "Deep learning in computed tomography pulmonary angiography imaging: A dual-pronged approach for pulmonary embolism detection," *Expert Syst. Appl.*, vol. 245, Jul. 2024, Art. no. 123029, doi: 10.1016/j.eswa.2023.123029.

MANSURA NAZNINE received the B.Sc. degree in computer science and engineering from RUET, in 2023. Fueled by a passion for cutting-edge technology, she envisions a future, where she will pursue advanced research in artificial intelligence and machine learning. Building on her strong foundation in image processing and data compression, she aims to explore novel applications in these fields. Her research interests encompass a diverse range of fields within computer science, with a primary focus on data compression, disease detection, and computer vision, particularly in the domain of classification. Her work in data compression reflects a commitment to developing efficient and innovative techniques to reduce the size of digital data while preserving its essential information. In the realm of disease detection, she explores cutting-edge technologies and methodologies to enhance early diagnosis and monitoring through computational approaches. Her future goals involve pursuing higher education to deepen her understanding and collaborating on interdisciplinary projects that align with her interest in the intersection of technology and environmental sustainability. Her ResearchGate profile: https://www.researchgate.net/profile/Mansura-Naznine.

ABDUS SALAM is currently pursuing the degree in electrical and computer engineering with Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh.

He is a Research Assistant with Qatar University Machine Learning Group, where he conducts research in his areas of interest. These include machine learning applications in domains, such as disease detection, classification, and segmentation, in medical images. His research is specifically centered toward creating advanced automated systems by image processing. His goal is to continue advancing his research skills and contribute innovative AI solutions to improve medical image analysis, food security, and agricultural sustainability on a global scale. With his demonstrated expertise and dedication to the field. He has co-authored a paper published in Frontiers in plant science. He is motivated by a strong desire to apply artificial intelligence to improve society. Furthermore, he is committed to improving current deep learning models, making them more relevant to various real-life situations, while also lowering their complexity and increasing their interpretability. His ResearchGate profile: https://www.researchgate.net/profile/Abdus-Salam-45.

MUHAMMAD MOHSIN KHAN received the master's degree in clinical research from Dresden University of Technology, Germany. He is currently a Neurosurgeon and a Clinical Researcher with the Hamad General Hospital, Qatar. He finished his training in Acgme-I accredited training program in Qatar. He received a certification in research from the T. H. Chan School of Public Health, Harvard University. He has done work on better understanding subarachnoid hemorrhage and is specializing in the integration of artificial intelligence (AI) into neurosurgical practice. His work focuses on enhancing and analyzing AI's impact on diagnostic accuracy, treatment planning, and patient management in neurosurgery and especially in subarachnoid hemorrhage. He has 35 plus peer-reviewed publications and is involved in many projects including Randomized controlled trials, and he continues to further expand the frontiers in neurosurgery and clinical research.

SADIA FARHANA NOBI graduated from Jahurul Islam Medical College in 2014. She is currently pursuing the Master of Public Health (MPH)degree in epidemiology with the Department of Public Health and Informatics, BSMMU. She completed her internship in 2015. She began her career as a Medical Officer at Ibn Sina Medical College Hospital. Since 2017, she has been serving as a Lecturer at Dr. Sirajul Islam Medical College. In addition, she is working as a Research Assistant with the Qatar University Machine Learning Group.

MUHAMMAD E. H. CHOWDHURY (Senior Member, IEEE) received the Ph.D. degree from the University of Nottingham, U.K., in 2014. He subsequently worked as a Postdoctoral Research Fellow with the Sir Peter Mansfield Imaging Centre, University of Nottingham. Currently, he is an Assistant Professor and the Program Coordinator of the Department of Electrical Engineering, Qatar University. He is a Prolific Researcher with several patents to his name, two edited books, and over 200 peer-reviewed journal articles, more than 30 conference papers, and multiple book chapters. His research interests include biomedical instrumentation, signal processing, wearable sensors, medical image analysis, machine learning, computer vision, embedded system design, and simultaneous EEG/fMRI. He is actively involved in leading several research projects funded by the Qatar Research, Development, and Innovation Council (QRDI) and internal grants from Qatar University, along with academic collaborations with HBKU and HMC. He has been recognized with several prestigious awards, including the COVID-19 Dataset Award, the AHS Award from HMC, and the National AI Competition Award for his contributions to the fight against COVID-19. His team earned a gold medal at the 13th International Invention Fair in the Middle East (IIFME). Additionally, he has been listed among the Top 2% of scientists in the world by Stanford University. He serves as an Associate Editor for *Computers and Electrical Engineering* and IEEE Access and a Topic Editor and a Review Editor for *Neuroscience* (Frontiers).