Development of an AI Model Based on Deep Learning for Detection, Classification, and Counting of Urinary Sediment in Microscopic Images

Jantira Paluka[1], Oanchitthaphon Ruthirawut[1]*
Aut Kongthong[1]
[1]Princess Chulabhorn Science High School Mukdahan, 281 Moo 6, Ban Nong Hoi, Bang Sai Yai Sub-district, Mueang Mukdahan District, Mukdahan 49000, Thailand
**E-mail: 05961@pccm.ac.th*

Abstract

Urine sediment examination is an essential process in medical diagnosis, as it aids in identifying diseases—particularly those of the kidney and urinary tract, or conditions accompanied by urinary changes—and constitutes a fundamental laboratory investigation. At present, urine sediment analysis still relies on microscopic examination by medical technologists, who must count and classify sediment types across multiple fields of a sample. This approach carries several limitations, including processing delays, inter-observer variability, and errors arising from human factors. In response to these challenges, and recognizing the advancement of artificial intelligence—especially deep learning models—the authors sought to develop a model capable of detecting and counting urine sediments from microscopic images. This was achieved by collecting and preparing a urine sediment image dataset, applying data augmentation, and annotating the location of sediments before feeding the data into a deep learning training pipeline using classification and object detection via the Roboflow Train 3.0 algorithm to localize and identify sediments, with output expressed as sediment type and count per field. The resulting model demonstrated effective performance: an optimal confidence threshold of 41% yielded an F1-score of 81.6%, with precision of 81.0% and recall of 83.1%, indicating a suitable balance between accuracy and detection capability. The mAP was approximately 79.4% and mAP@50:95 approximately 56%. Per-class analysis showed high accuracy for RBC and WBC, while low-data classes such as bacteria and yeast exhibited higher misclassification rates due to class imbalance and morphological similarity. Overall, the model shows strong potential for detecting, classifying, and counting urine sediments, and can be further developed in future work

keywords : Urine Sediment, Medical AI, Deep Learning, Object Detection, Microscopic Image

Introduction

Urine sediment examination is a fundamental screening investigation in the clinical laboratory, providing evidence of kidney and urinary tract disease and of conditions manifesting through changes in the urine; combined with physical and chemical examination, it constitutes a complete routine urinalysis. Sediment examination remains manual, performed by medical technologists using light microscopy to count and classify elements across multiple fields per specimen. This yields processing delays, inter-observer variability, and error arising from human factors. Deep learning object detection offers a route to addressing these limitations. This study therefore aims to develop and evaluate a model that detects, classifies, and counts urine sediments from microscopic images, reporting type and count per field — as a decision-support tool for high-volume urinalysis and a prototype for broader laboratory application.

Methodology

1) Materials and study sites

Model development was performed on an ASUS Vivobook 15 laptop. Urine sediment images were captured using a Samsung Galaxy S23 FE. Work was conducted at Princess Chulabhorn Science High School Mukdahan (Mukdahan Province), Fort Phrayodmuangkhwang Hospital (Nakhon Phanom Province), and Nakhon Phanom Hospital (Nakhon Phanom Province).

2) Image acquisition and dataset assembly

Microscopic images containing urine sediments of various types were compiled from two sources: clinical laboratories processing authentic patient urine specimens, and publicly available online image databases of urine sediment. Images were collected under a range of acquisition conditions — including variation in illumination intensity, stain colour, and lens quality — in order to maximize dataset diversity and approximate real-world conditions as closely as possible. All images were uploaded to a project created on the Roboflow platform.

3) Class definition

The sediment types of interest were defined *a priori*, selected on the basis of frequency of occurrence and diagnostic significance. The model was trained to classify detected objects into the following seven classes: Red blood cells (RBC), White blood cells (WBC). Epithelial cells, Crystals — including calcium oxalate, uric acid, and triple phosphate (struvite), Bacteria, Yeast, Casts — including hyaline and granular casts

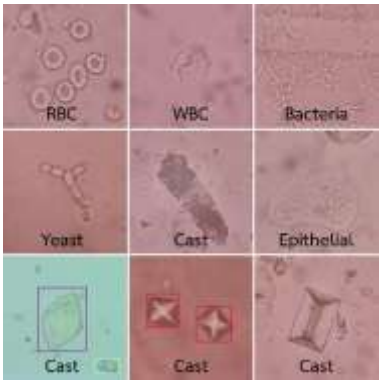

Fig. 1 Image of each urine sediment class

These classes cover the majority of sediments encountered in general patient urine. Explicit class definition allows the model

to correctly distinguish each sediment type detected, and prepares it to learn the morphological differences between them.

4) Annotation and dataset partitioning

Following image selection and preparation, each image was annotated to establish the reference dataset from which the model would learn, specifying the position and type of every sediment particle present. Annotation proceeded as follows: a bounding box was drawn around each visible sediment object; a class label was assigned to that box; and the image was marked as annotated. This was repeated until every image in the folder was complete. Annotations were then verified for accuracy before the images were compiled into the dataset, which was partitioned into a training set (70%), a validation set (20%), and a test set (10%).

5) Data augmentation

Augmentation was applied to expand the dataset, comprising brightness adjustment of ±15% and exposure adjustment of ±10%.

6) Model architecture and training

The Roboflow 3.0 object detection model was selected. It is offered on the Roboflow platform, is compatible with YOLOv8, and offers a balance between inference speed and accuracy. Of the five available model sizes, the *Accurate* configuration was used.

7) Performance evaluation

Following training, model performance was assessed on the held-out test set — urine sediment images not seen by the model during training — in order to determine how well it generalized to novel data. Evaluation was based on the following standard metrics: precision, recall, accuracy, F1-score, mean average precision (mAP), and the confusion matrix.

Result and Discussion

1) Training dynamics

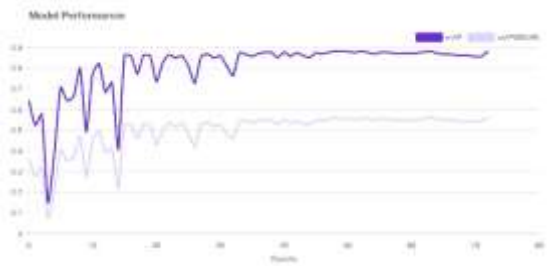

Fig. 2 Model performance curves

Fig. 3 Box loss, class loss, and object loss

Figures 2 and 3 show the learning trend across training epochs. All loss components decreased rapidly during the early phase and gradually stabilized after approximately epoch 30–40. Box loss decreased continuously, indicating progressive improvement in bounding-box localization. Class loss decreased markedly, reflecting increasingly accurate class discrimination. Object loss declined during the early epochs but rose slightly toward the end, which may reflect a trade-off between objectness confidence and classification as the model approached convergence. The loss curves showed no severe divergence, indicating stable training

2) Training–validation convergence

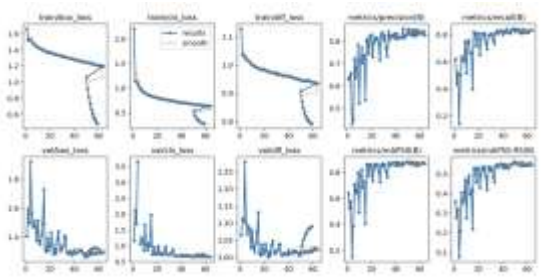

Fig. 4 train/val box_loss, cls_loss, dfl_loss; metrics/precision(B) and metrics/recall(B)

The training and validation curves for box_loss, cls_loss, and dfl_loss decreased in parallel without a pronounced gap, indicating that the model did not exhibit significant overfitting.

The precision(B) and recall(B) curves increased steadily. Fluctuation in the early epochs is characteristic of object detection training, particularly on datasets exhibiting class imbalance.

The mAP50(B) and mAP50-95(B) curves increased and plateaued toward the end of training, reflecting convergence. Final values were 79.4% for mAP@50 and 56.0% for mAP@50:95. Under the criterion that a detection is counted as correct only where it overlaps the ground-truth position by at least 50% (IoU ≥ 0.50), the model therefore localized and classified urine sediments correctly in approximately four of five cases, averaged across all classes.

3) Performance by class

Fig. 5 Performance by class

Figure 5 disaggregates correct predictions, misclassifications, false negatives, and false positives by class.

RBC recorded the highest count of correct predictions, consistent with its large number of ground-truth instances and relatively distinct morphology. It also accumulated the most false positives, however — predominantly debris and artifacts of similar size and refractility — which depressed its class-level precision. WBC and crystal combined high correct-prediction counts with few errors of either type, yielding balanced precision and recall, and were the classes on which the model performed most reliably.

Bacteria and yeast, the two classes least represented in the training data, showed markedly higher proportions of false negatives and misclassifications. Two factors plausibly contribute: class imbalance, both being represented by an order of magnitude fewer instances than RBC; and low feature distinctiveness, both being small, low-contrast, and morphologically similar to background debris, such that even a balanced dataset would present a harder discrimination problem.

Performance therefore varied with both the quantity of data available per class and the visual distinctiveness of that class's features — a limitation common to multi-class object detection.

4) Confidence threshold optimization

Fig. 6 Relationship between confidence threshold and precision, recall, and F1-score

F1-score peaked at 81.6% at a confidence threshold of 41%, balancing precision (81.0%) against recall (83.1%). Below this threshold, recall increased at the cost of precision as false positives multiplied; above it, precision gained only marginally while recall fell substantially, causing F1 to decline rapidly.

The appropriate operating threshold depends on clinical role, not mathematics alone. Deployed as a decision-support aid — the application envisaged here — the system's outputs are reviewed by a medical technologist, and a missed sediment element that is never re-examined carries greater consequence than a flagged artifact dismissed on review. This asymmetry argues for operating at or slightly below the F1-optimal threshold, favouring recall. A threshold of 41% is therefore appropriate for the present use case.

5) Confusion matrix

Fig. 7 Confusion matrix

The confusion matrix confirms high true-positive counts for RBC, WBC, and epithelial cells. Misclassification persisted between morphologically similar classes, notably epithelial cells versus casts and versus crystals — consistent with their underlying visual similarity, as hyaline casts and squamous epithelial cells share low refractility and comparable dimensions at the magnification used, leaving the model to discriminate on features that are marginal at this image resolution. False positives were concentrated in RBC, likely arising from debris resembling erythrocytes.

These results reflect a limitation of feature-level discrimination and indicate two directions for improvement: greater dataset diversity, particularly for under-represented classes, and more advanced augmentation than the brightness and exposure adjustments applied here.

6) Comparison with prior work

The mAP@50 of 79.4% obtained here falls below that of comparable urine sediment detection systems: Liang et al. (2018) reported 86.9% using their DFPN architecture, and Suhail and Brindha (2024) reported 85.8% using YOLOv5l with genetic algorithm hyperparameter optimization — placing the present model within six to eight percentage points of both. The gap is attributable to dataset scale, the absence of hyperparameter optimization, and the use of a general-purpose training platform rather than an architecture modified for urine sediment images. Comparison should be treated with caution, as these studies employ different datasets and

class definitions; mAP is not directly transferable across benchmarks.

The per-class pattern observed here is nonetheless consistent with the literature. Liang et al. identified class confusion as the principal obstacle in urine sediment detection, and the architectural modifications proposed since — attention mechanisms, feature pyramid enhancement, deformable convolution (Zhang et al., 2025) — target the same difficulty in discriminating morphologically similar, low-contrast particles that this model encountered with bacteria, yeast, and epithelial/cast confusion. That a general-purpose detector on a modest dataset reproduces the field's known failure modes at performance within six to eight percentage points of purpose-built architectures supports the feasibility of the approach while indicating where refinement is required.

Conclusions

A deep learning object detection model was developed for the detection, classification, and counting of urine sediments in microscopic images. The model achieved an mAP@50 of 79.4% across all classes under an IoU threshold of 0.50, with mAP@50:95 of 56.0%. At the optimal confidence threshold of 41%, precision, recall, and F1-score were 81.0%, 83.1%, and 81.6% respectively.

Performance varied by class. WBC and crystal were detected most reliably, combining high correct-prediction counts with balanced precision and recall. RBC yielded the highest absolute number of correct detections but also the highest number of false positives, reducing its class-level precision. Bacteria and yeast performed poorest, attributable to their limited representation in the training data together with their low visual distinctiveness against background debris.

These results indicate that a general-purpose object detection model, trained on a modestly sized dataset of urine sediment images, can localize, classify, and count sediment elements at a level of accuracy approaching that of purpose-built architectures reported in the literature. Principal limitations are the size and class balance of the dataset and the absence of hyperparameter optimization; both are addressable. Future work should prioritize expanding the representation of under-represented classes, evaluating performance on hospital-acquired images separately from public-dataset images, and validating counts per high-power field against medical technologist reference counts on the same specimens.

Acknowledgments

The authors thank Mr. Aut Kongthong, Science and Technology Department, Princess Chulabhorn Science High School Mukdahan, for his guidance throughout this project; Miss Inthira Tussakorn (Nakhon Phanom Hospital) and Mr. Traipop Panchanen (Fort Phrayodmuangkhwang Hospital), medical technologists, together with the staff of their respective Laboratory and Pathology Departments, for their instruction and for preparing the urine specimens; Mr. Sak Rungsaeng, Director of Princess Chulabhorn Science High School Mukdahan, and the

teachers of the Science and Technology Department for their support; and our parents and guardians for their encouragement.

References

Aleman, J. S., & Reyes-Duke, A. M. (2025). Performance analysis of Roboflow 3.0 and YOLO-NAS in PCB defect detection compared to YOLOv9 and RT-DETR using augmented image dataset. In *Proceedings of the 10th International Conference on Control and Robotics Engineering (ICCRE)* (pp. 229–234). https://doi.org/10.1109/ICCRE65455.2025.11093496

Jocher, G., & Qiu, J. (2024). *Explore Ultralytics YOLOv8*. Ultralytics YOLO Docs. Retrieved December 25, 2025, from https://docs.ultralytics.com/models/yolov8/

Liang, Y., Tang, Z., Yan, M., & Liu, J. (2018). Object detection based on deep learning for urine sediment examination. *Biocybernetics and Biomedical Engineering, 38*(3), 661–670. https://doi.org/10.1016/j.bbe.2018.06.003

Manski, D. (n.d.). *Urine analysis: Sediment and dipstick examination*. Urology Textbook: Clinical Essentials. Retrieved September 10, 2025, from https://www.urology-textbook.com/urine-analysis.html

Mhzr. (2024). *Urine dataset* [Data set]. Roboflow Universe. Retrieved December 26, 2025, from https://universe.roboflow.com/mhzr-lm23t/urine-l7c1h

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports, 14*, Article 6086. https://doi.org/10.1038/s41598-024-56706-x

Suhail, K., & Brindha, D. (2024). Microscopic urinary particle detection by different YOLOv5 models with evolutionary genetic algorithm based hyperparameter optimization. *Computers in Biology and Medicine, 169*, Article 107895. https://doi.org/10.1016/j.compbiomed.2023.107895

Yan, M. (2020). *Urinary sediment dataset* [Data set]. GitHub. Retrieved December 26, 2025, from https://github.com/174614361/Urinary-Sediment-Dataset

Zhang, S., Bao, X., Wang, Y., & Lin, F. (2025). Urine sediment detection algorithm based on channel enhancement and deformable convolution. *Journal of Imaging Informatics in Medicine, 38*(4), 2355–2366. https://doi.org/10.1007/s10278-024-01321-5