KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Stylianos G Mouslech, Sven Wijnants, Anne-Lisanne van der Schagt, Lieve Van Hoovels, Roxane Deley, Matthijs Oyaert, Jana Neirinck, Jaak Billen, Glynis Frans, Maarten De Vos, Towards Clinical Integration of Deep Learning–Based Classification of Urinary Sediment Particles from Digital Microscopy Images: A Prospective Study, *Clinical Chemistry*, 2026;, hvaf182
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://doi.org/10.1093/clinchem/hvaf182
Journal homepage	https://academic.oup.com/clinchem
Author contact	stylianos.mouslech@kuleuven.be
Abstract	**Background** Urinalysis is a standard clinical test that includes the microscopic examination of urinary sediment to identify formed elements. Manual evaluation by laboratory technicians is time-intensive and subject to human error. Automated analysis using digital microscopy images presents a potential alternative. This study evaluates the integration of a deep learning approach to automatically classify urinary sediment images in the clinical laboratory, including independent prospective validation of its performance. **Methods** An annotated data set comprising 13 classes of urinary sediment elements was created from a database of Sysmex UD-10 digital microscope images. An EfficientNet-based model was trained and tested

across three experimental scenarios to evaluate the effects of data collection strategies on performance. Uncertainty calibration was examined. The model's robustness and interpretability were examined using gradient-weighted class activation mapping (Grad-CAM) to visualize influential image regions and t-distributed stochastic neighbor embedding (t-SNE) to analyze learned feature embeddings. Lastly, a graphical user interface was developed for a prospective evaluation in the laboratory.

Results

The model achieved approximately 97% overall accuracy on the test set. Experiments revealed sensitivity to data set variability, suggesting that performance may improve by integrating additional training examples. Confidence scores aligned with accuracy, and interpretability analyses showed that the model focused on relevant image regions and learned embeddings demonstrated clear class separation. In the prospective evaluation, top 1 and top 3 accuracies decreased to approximately 78% and 92%, respectively.

Conclusions

Our results indicate that a lightweight deep learning model can achieve high performance in classifying urine particles. Analysis of discrepancies between retrospective and prospective evaluations provides important insights toward reliable clinical application.

IR

(article begins on next page)

TITLE PAGE

Manuscript Title: Towards Clinical Integration of Deep Learning–Based Classification of Urinary Sediment Particles from Digital Microscopy Images: A Prospective Study

Running header title: DL-Based Urinary Sediment Analysis

Authors' Full Names: Stylianos G. Mouslech[1], Sven Wijnants[2], Anne-Lisanne van der Schagt[2], Lieve Van Hoovels[3,4], Roxane Deley[4] , Matthijs Oyaert[5], Jana Neirinck[5] , Jaak Billen[2], Glynis Frans[2,3*], Maarten De Vos[1,6*]

Authors' Affiliations:

[1] STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics—Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

[2] Department of Laboratory Medicine, UZ Leuven, Leuven, Belgium

[3] Clinical and Diagnostic Immunology, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium

[4] Department of Laboratory Medicine, AZORG Hospital Aalst, Aalst, Belgium

[5]Department of Laboratory Medicine, University Hospital Ghent, Ghent, Belgium.

[6] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

* Both senior authors contributed equally to this study.

Corresponding Author:

Full Name: Stylianos Georgios Mouslech

Adress: Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

E-mail: stylianos.mouslech@kuleuven.be

Telephone: +32469433198

Keywords: Urinary Sediment Analysis, Deep Learning, Digital Microscopy, Artificial Intelligence in Laboratory Medicine, Prospective Evaluation, Urinalysis Automation

Abstract word count: 249 (Including structured headers)

Manuscript word count: 3497

Number of table(s): 2

Number of figure(s): 4

List of Abbreviations:

AI — Artificial Intelligence

ML — Machine Learning

DL — Deep Learning

CNN — Convolutional Neural Network

mAP — Mean Average Precision

TP — True Positive

TN — True Negative

FP — False Positive

ECE — Expected Calibration Error

Grad-CAM — Gradient-weighted Class Activation Mapping

t-SNE — t-Distributed Stochastic Neighbor Embedding

GUI — Graphical User Interface

ECDF — Empirical Cumulative Distribution Function

KS — Kolmogorov-Smirnov

WBC — White Blood Cell

NSEC — Non-Squamous Epithelial Cell

IFU — Instructions for Use

ABSTRACT

BACKGROUND:

Urinalysis is a standard clinical test that includes the microscopic examination of urinary sediment to identify formed elements. Manual evaluation by laboratory technicians is time-intensive and subject to human error. Automated analysis using digital microscopy images presents a potential alternative. This study evaluates the integration of a deep learning approach to automatically classify urinary sediment images in the clinical laboratory, including independent prospective validation of its performance.

METHODS:

An annotated dataset comprising thirteen classes of urinary sediment elements was created from a database of Sysmex UD-10 digital microscope images. An EfficientNet-based model was trained and tested across three experimental scenarios to evaluate the effects of data collection strategies on performance. Uncertainty calibration was examined. The model's robustness and interpretability were examined using Grad-CAM to visualise influential image regions and t-SNE to analyse learned feature embeddings. Lastly, a graphical user interface was developed for a prospective evaluation in the laboratory.

RESULTS:

The model achieved approximately 97% overall accuracy on the test set. Experiments revealed sensitivity to dataset variability, suggesting that performance may improve by integrating additional training examples. Confidence scores aligned with accuracy, and

interpretability analyses showed that the model focused on relevant image regions and learned embeddings demonstrated clear class separation. In the prospective evaluation, top-1 and top-3 accuracies decreased to approximately 78% and 92%, respectively.

CONCLUSIONS:

Our results indicate that a lightweight deep-learning model can achieve high performance in classifying urine particles. Analysis of discrepancies between retrospective and prospective evaluations provides important insights towards reliable clinical application.

TEXT

INTRODUCTION

Urinalysis is one of the most requested clinical tests in laboratory medicine (1) . An integral part of urinalysis is the identification and quantification of formed elements in urine sediment, such as cells, urinary casts, crystals, and microorganisms. Traditionally, this analysis is performed manually by trained laboratory technicians who examine urine samples under a microscope. However, manual analysis is time-consuming, subject to human error, and requires thorough training of a lab technician (1). Over the last few years, new technological advancements have led to the introduction of automation in urine sediment analysis (2). Most automated analysers rely on two different techniques: (fluorescence) flow cytometry and automated imaging technologies, which utilise a digital microscope with image processing software to identify and quantify various urine particles according to their dimension (2,3). The use of automation in urine screening devices can make urine sediment analysis more economically viable and aid in reducing examination errors (1).

At the same time, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has shown great potential in medical diagnostics, including laboratory medicine

(4,5). In the context of urinalysis, AI can aid in identifying and classifying formed elements in digital microscopy images of urine samples. Initial attempts to automate the classification of urinary sediment relied on traditional machine-learning methods. They demonstrated promising accuracy but were limited in the number of particle classes they could recognise. (6–8).

Recent advances in deep learning have enabled various automated approaches to urinary sediment analysis. These approaches differ in network architecture, feature extraction strategies, and the number and type of labels they provide. Some approaches have formulated the task as an end-to-end object detection task or as a classification task on patches of the images, mainly containing a single class of particles. For example, Dong et al. (9) proposed a modified YoloV3 (10) object detection algorithm that can obtain the best mean average precision (mAP) of 90.1% in categorising five types of urinary sediment particles in their collected dataset. Similarly, Chen et al. (11) proposed a specialised variant of YOLOv5 that was evaluated on two datasets of eight predefined categories, achieving mAP of 91.7% and 68.40%, respectively. When formulated as a classification problem, several transfer learning studies leveraging pre-trained models have been proposed and compared (12–16). A notable example is the work by Ji et al. (17) where they developed a custom Convolutional Neural Network (CNN) architecture that was trained using a large dataset of 10 categories and achieved a test accuracy of 97%. In comparison, we propose a single, lightweight end-to-end CNN model to classify thirteen urinary sediment classes, a larger set of classes than most existing methods, which frequently rely on multiple model sub-components.

While substantial progress has been made in achieving highly automated urinary sediment analysis performance, the variability in data sources, particle classes, and metrics complicates performance comparisons across studies. These challenges highlight the necessity for a solution that not only achieves high performance but also explores issues of data variability, algorithm explainability and effectiveness in a real-world clinical laboratory (2). To address these

challenges, this study aimed to develop a clinically applicable tool based on a lightweight convolutional neural network for accurate classification of urinary sediment images derived from a digital microscope, assess the model's robustness and interpretability, and evaluate its suitability for integration into routine clinical laboratory workflows.

Materials and Methods

Data Set Curation

Data was collected at the Department of Laboratory Medicine of UZ Leuven in two stages between 6-6-2023 and 31-5-2024. The data collection and study were approved by the local Ethics Committee of UZ Leuven (study number S69284). First, a large dataset (Dataset A) was curated from random, anonymised urine samples. Images were labelled by consensus among experienced urinalysis lab technicians. The images, captured at varying resolutions using the Sysmex UD-10 digital microscope (Sysmex, Kobe, Japan), contain segmented patches of particles from the samples, which are not automatically classified. This dataset primarily included images of particles considered easier to identify, representing textbook examples without technician disagreement. This resulted in 40,794 labelled images across 13 classes. Subsequently, a second dataset (Dataset B) was collected, focusing on images that presented more significant, real-world labelling challenges (such as technician disagreements or particles with degradations), yielding an additional 8,463 images. The particle classes reflect the specifications and diagnostic needs of the UZ Leuven laboratory, with the number of images per class detailed in Table 1. This two-stage approach in data collection allowed the experimental design outlined below.

For the prospective evaluation of the trained model under real-world conditions, a third set of data was collected through integration of the classification tool into routine laboratory workflow. A lab technician randomly selected new image data not used in previous training and testing and

uploaded them into the application. This included data from samples that were chosen completely randomly without any specific inclusion or exclusion criteria. To better evaluate the practical performance, the technician corrected the model or tagged examples as “unknown” when the technician was uncertain about classifying the particle.

A final external dataset of images, obtained from 45 anonymized samples analyzed using an additional UD-10 microscope (a demonstration device owned by Sysmex Benelux), was independently labeled by two other clinical centers, AZORG Hospital Aalst and Ghent University Hospital (UZ Gent), to assess the model’s robustness in a different acquisition and annotation settings. This resulted in two external datasets, namely Dataset E1 & E2, respectively. More details are included in Supplemental Method A.

Experiment Design

The experiments were proposed to evaluate the model’s performance by utilising the two described datasets. This experiment design differs from the standard approach in related literature, where a single, static batch of data is typically divided into training, validation, and test splits. The experiments aimed to investigate the impact of incremental data collection on generalizability.

Both datasets were divided into training, validation, and test sets, with a 90/10 training/test split and an 80/20 train/validation split, resulting in a 72/18/10 distribution for training, validation, and testing. Class distributions were maintained across each split to ensure consistency in representation. Three experimental scenarios were devised to assess this setup: a baseline evaluation where the model was trained, validated, and tested on the larger Dataset A, establishing initial performance metrics (Scenario 1); a cross-dataset evaluation, where the model trained on Dataset A was tested on Dataset B to evaluate generalisation (Scenario 2); and a combined dataset evaluation (Scenario 3), in which both datasets were merged and re-split into

training, validation, and test sets to evaluate performance on a more comprehensive dataset (see Figure 1). Each scenario was conducted as a standalone experiment with an independent training pipeline, and the test set for each scenario was held out and evaluated once to avoid information leakage. The best performing model was then used for analyzing the model's reliability and the prospective evaluation.

Model Development

The EfficientNet architecture, proposed by Tan and Le (18), was selected as the main convolutional network model due to its demonstrated performance and computational efficiency, making it suitable for application within a laboratory context.

To adapt EfficientNet for this study, a new classification head with 13 neurons (corresponding to each particle class) was appended to the architecture. The model's convolutional layers were initialised with ImageNet (19) pre-trained weights, a common approach in data-limited settings to leverage prior knowledge and accelerate convergence. Image data augmentation techniques were applied during training to mitigate overfitting. Each training example was randomly flipped horizontally and vertically with 50% probability and rotated in a range of 0 to 60 degrees. These augmentations were intended to improve generalisation by assisting the model in learning representations that are invariant to orientation changes. This approach is particularly beneficial for classes with fewer samples (20). Lastly, all images were resized to 224×224 to match the EfficientNet-B0 input size. Further details on the model and training are provided in Supplemental Method B.

A series of ablation studies were also performed to evaluate the contribution of individual components, such as data augmentation strategies and architectural modifications, to overall performance. The results are discussed in Supplemental Method C.

Model Evaluation

The evaluation focused on two aspects: performance and reliability. Performance was assessed using standard classification metrics: accuracy, precision, recall, and F1-score, which are commonly reported in related literature. To account for the multi-class setting, macro-averaged metrics were computed. Formal definitions of all metrics are provided in the Supplemental Method D.

To assess reliability, we further examined the model's output calibration (21), embedding space, and interpretability. Regarding calibration, the model produces a probability distribution over the classes. Calibration evaluates whether predicted probabilities correspond to true likelihoods. The Expected Calibration Error (ECE) metric approximates the true calibration error (22). Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) (23) was used to highlight image regions the model "focuses" on, while t-SNE (24) was applied to visualize the high-dimensional feature embeddings. The combination of these techniques was chosen to produce a deeper qualitative evaluation of the model.

GUI Application and Prospective Evaluation

As part of the work, the best-performing model from the previous evaluation was used to develop a Graphical User Interface (GUI) application to aid technicians in classifying urinary sediment images from the Sysmex UD-10 digital microscope. The local GUI was implemented in the Python programming language. PySide6 (25) was used for the design and development of the user interface, enabling integration and execution of the model produced as part of this research. Technical details and the source code are publicly available at: https://github.com/SteliosMouslech/UrineAI. The application was trialed in the clinical laboratory of UZ Leuven to explore how the optimised model performs in real-world use. Integrating machine learning models into clinical software is still a developing and challenging field (26), with best

practices still being developed. In this case, the design was guided by the user requirements, desired functionality, and by considering its integration into the current urinalysis workflow at the Department of Laboratory Medicine of UZ Leuven.

The application allows technicians to input data, run the model on UD-10 microscope images, and visualise the results. Technicians can also modify model predictions, enabling evaluation of its real-world performance. After implementation and training, the usability of the application was assessed using a usability questionnaire described in the Supplemental Method E.

For the prospective evaluation, model embeddings were visualized with t-SNE. Empirical cumulative distribution functions (ECDFs) were computed for correct, incorrect, and unknown predictions to assess output distribution differences. To estimate uncertainty, 95% bootstrap confidence intervals were calculated from 1000 resamples per group. Finally, distributional differences were tested for significance using the two-sample Kolmogorov-Smirnov test.

RESULTS

Evaluation Of Different Experimental Scenarios

In the first experimental scenario, which involved training and testing on Dataset A, the model achieved overall accuracy and F1 scores of 0.96 and 0.93, respectively. In the second scenario, where the model from the first scenario was applied to Dataset B, performance was 0.93 and 0.90, respectively.

After combining both datasets in the third scenario, the model achieved its highest overall performance, reaching an accuracy of 0.97 and an F1 score of 0.95. The detailed results are presented in Table 2. Given this improvement, the remainder of the analysis focused on the model trained in the third scenario, which incorporates the entire dataset.

Class-Specific Results

Class-specific performance varied, with the model achieving high performance in most classes. However, a trend emerged where lower F1 scores were achieved in more underrepresented classes. This outcome aligns with expectations, as data imbalance is known to impact model performance in underrepresented classes (27). For example, the model achieved an F1 score of 0.85 on white blood cell (WBC) clumps (n = 57 images) and an F1 score of 0.9 on the mucus class (n = 52 images). The highest F1 scores of 0.99 were observed in the classes of yeasts (n = 427 images) and bacteria (n = 583 images) (Table 2).

Evaluating Reliability

The model produced highly confident predictions, with most top predictions exceeding a confidence threshold of 0.9, as shown in Figure 2. A low ECE of approximately 0.01 was observed, suggesting effective calibration of predicted probabilities. However, as a significant proportion of predictions fall within the highest confidence range of 0.9–1.0, this heavily influences the weighted average, potentially leading to an “overconfident” model. The exact values of accuracy and average confidence for each confidence bin is provided in Supplemental Table 1.

Examining the areas of the image that the model focuses on when making predictions provided useful information. These visualisations indicate that the model consistently focuses on the relevant particles within the images rather than background artefacts. One type of error was found in images with multiple particle types present in the same image. In that case, when probed on each of the present classes, the model highlighted the corresponding particle in the image, further enforcing the idea that the model focuses on meaningful parts of the image (see Figure 3).

Lastly, visualising the embeddings revealed that the model seems to capture representations that show low inter-class similarity and low intra-class variation (see Figure 4).

Furthermore, the relative distances between clusters indicate a possible relationship with the physical characteristics of the different particles.

Prospective Evaluation Results

The prospective evaluation resulted in 4544 labelled images and 3371 labelled as unknown. The prevalence of different categories in the labelled images differs from the previous datasets and is displayed in Supplemental Table 2. A significant drop in performance compared to the test data was observed. Overall accuracy was 78% and F-1 Score to 0.66. The performance drop was mitigated when assessing if the correct answer was within the top 3 predictions of the model (92%). A similar trend was noticed in class-specific performance (Supplemental Table 2).

To examine performance discrepancies, we re-applied t-SNE on all model embeddings from the prospective evaluation and original test data. Two key findings emerged. First, most classes remained aligned with their original clusters, retaining the model's learned structure. Second, new clusters appeared between existing ones, often linked to misclassifications (see Supplemental Figure 1). Visualizing only misclassified samples confirmed these intermediate clusters represented errors (Supplemental Figure 2). For instance, WBC clumps shifted closer to WBC, amorphous precipitates spread towards bacteria, and muci, pathological casts, and non-squamous epithelial cells (NSEC) embeddings appeared more intertwined. Such overlaps explain the model's gains in top-3 predictions.

Lastly, we noticed that the model's confidence levels differed in three situations: when the user was unsure of the label (Unknown), when the model's prediction matched the user's choice (Correct), and when they disagreed (Incorrect). This suggests that the confidence score could help flag cases where the model's answer might be less reliable and needs more attention from the user. ECDFs were used to visualize these distributions (Supplemental Figure 3). The mean confidence values differed across groups (Correct = 0.95, Incorrect = 0.79, Unknown = 0.71), and

the ECDFs revealed that incorrect predictions and unknown data both exhibited a heavier tail towards lower confidence scores, indicating increased model uncertainty. Statistical comparison using the two- sample Kolmogorov-Smirnov test confirmed significant distributional differences between correct and incorrect predictions ($p < 0.001$), as well as between correct and unknown samples ($p < 0.001$). These findings suggest that prediction confidence could serve as a useful indicator of reliability and potential model uncertainty in real-world implementation of such an application in the lab.

Besides analytical reliability, the application usability was also assessed (Supplemental Table 3). All ten laboratory technicians felt confident using the application after training and reading the Instructions for Use (IFU), though 3/10 indicated the need for extra help initially. All users could interpret the results. The median usability score was 4/5 (range 2–5). Half would use the application in its current state, the others after adaptations. Suggested improvements mainly concerned visualisation (e.g., selective classification, grouped labels) and workflow (clearer clinical guidelines on software usage).

External Evaluation Results

In the twice-independently labelled external datasets, results were comparable to the prospective evaluation (top-1 accuracy 0.78 vs. 0.81, top-3 accuracy 0.90 vs. 0.92, macro F1: 0.58 vs. 0.48). Macro-averaged F1, precision, and recall were lower and more variable, likely due to some classes having few examples. In contrast, weighted F1-scores remained high (0.82 and 0.85). These findings highlight potential discrepancies from annotator biases and differences in labelling protocols across sites, further detailed in the Supplemental Table 4.

DISCUSSION

This study investigated the feasibility of automating the classification of urinary sediment particles in digital microscopy images, emphasizing clinical applicability. The main contribution is the

development of a lightweight convolutional neural network based on EfficientNet, trained on a collected dataset of urinary sediment images with a large set of different classes. We also assessed data collection strategies and their effects on performance. We focused on assessing the model's robustness for clinical integration by designing multiple training scenarios and proposing a multi-layered evaluation including an external evaluation of the trained model.

A key contribution is the prospective evaluation of the developed tool within a real clinical setting. Beyond algorithm development, the work included the integration of the trained model into a usable software application, integrated into the daily urinalysis workflow of laboratory technicians in a trial setting. The tool provided top-1 and top-3 predictions alongside confidence scores, visible to staff during their routine review of urinary sediment images. These features were designed to assist decision-making and showcased the potential to reduce manual workload and improve diagnostic consistency. Preliminary feedback from technicians highlighted both potential benefits and areas for improvement.

Our results suggest that a fine-tuned convolutional neural network can effectively classify urinary sediment images. While direct performance comparisons with existing studies are difficult due to variations in datasets and evaluation protocols, our results fall within a similar performance range, even when classifying a larger set of particle categories (n = 13) than most related literature. The trained model learned meaningful feature representations, and our comparison of different data curation strategies showed that incorporating more diverse and difficult examples improved overall performance. On the other hand, highly curated datasets can lead to overly optimistic performance estimates.

When transitioning from retrospective testing to real-world use, a performance drop was observed. However, we show significant differences in model output distributions in the cases where the model is correct, incorrect, or faced examples unknown to the technicians., While the performance gap is mostly mitigated when accounting for the top-3 predictions of the model, it

highlights common challenges in clinical AI implementation and may reflect data drift (28–30), or annotation variability between evaluations. This drop in performance is also consistent with the external validation results, which further showcase the potential annotation variability in different centers. In general, and as discussed above, some errors arise from patches containing multiple particle types (Figure 3) but only a single label, or from labelling variability. Additional contributors include class imbalance, morphological resemblance across classes, and artefacts.

This leads us to the study's limitations. One limitation is that we did not investigate strategies to address the performance decline observed during the prospective evaluation. Such a drop is expected, as retrospective datasets are typically more curated. However, the presence of “unknown” labels and shifts in model output distributions indicate that some inputs fall outside the training domain. Methods to detect and manage these cases could improve robustness. For example, generative models can identify atypical samples and provide interpretable uncertainty estimates to support clinical review (31), while active learning frameworks that incorporate uncertainty estimation throughout development may further enhance performance (32). Future work should explore these strategies in urinalysis workflows and additionally assess tool effectiveness in terms of turnaround time, workload reduction, and usability.

Another limitation is that all training data were acquired on a single digital microscope (Sysmex UD-10) under uniform operating conditions. While all UD-10 devices use identical reagents, maintenance protocols, and quality control materials, suggesting consistent optical settings across UD-10 devices, this limits generalization to other types of microscopes. The study also did not thoroughly assess inter-annotator or inter-laboratory agreement, as the same technicians performed all labelling, potentially propagating bias. Future work should involve multiple experts from different centers, blinded multi-centre datasets, and diverse imaging devices and environments. Factors such as microscope type (manual vs. automatic, brightfield vs. phase contrast) and image representation (whole images vs. cropped particles) strongly

influence preprocessing and model design, challenging transferability across settings. Since integrating more and more complex examples can aid overall performance, continual learning approaches might be worth exploring (33). Both data- and model-centric strategies will be essential for improving generalization and robustness toward reliable clinical tools (34).

In summary, this study presents a step toward the clinical integration of deep learning for urinary sediment analysis. By combining model development with prospective validation in a laboratory setting, we demonstrate the integration of an AI-assisted tool to support routine urinalysis workflows. Our approach addresses the important need for evaluating such systems in real-world conditions, an aspect that is often underrepresented in similar research. With our findings, we highlight both the potential and the challenges of clinical deployment. Addressing those challenges will be key to developing a useful tool for routine laboratory practice.

DECLARATION OF CONFLICTING INTERESTS

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank Sysmex Benelux (Nancy Van Bijlen) for providing images from their UD-10 demonstration device. This work was supported by FWO Research Project G046925N

REFERENCES

1. Zaman Z. Automated urine screening devices make urine sediment microscopy in diagnostic laboratories economically viable. Clinical Chemistry and Laboratory Medicine (CCLM). 2015;53:s1509–11.

2. De Bruyne S, De Kesel P, Oyaert M. Applications of artificial intelligence in urinalysis: Is the future already here? Clin Chem. 2023;69:1348–60.

3. Kouri TT, Hofmann W, Falbo R, Oyaert M, Schubert S, Gertsen JB, et al. The EFLM European Urinalysis Guideline 2023. Clin Chem Lab Med. 2024;62:1653–786.

4. Khalifa M, Albadawy M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. Computer Methods and Programs in Biomedicine Update. 2024;5:100146.

5. Jackups R. Deep Learning Makes Its Way to the Clinical Laboratory. Clin Chem. 2017;63:1790–1.

6. Dong L, Yuan S, Liu G, Zhou L, Li Y. Classification of Urinary Sediments Image Based on Bayesian Classifier. 2007 International Conference on Mechatronics and Automation. 2007. page 556–60.

7. Zhou Y, Zhou H. Automatic Classification and Recognition of Particles in Urinary Sediment Images. In: He X, Hua E, Lin Y, Liu X, editors. Computer, Informatics, Cybernetics and Applications. Dordrecht: Springer Netherlands; 2012. page 1071–8.

8. Li C, Tang YY, Luo H, Zheng X. Join Gabor and scattering transform for urine sediment particle texture analysis. 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF). 2015. page 410–5.

9. Dong S, Zhang S, Jiao L, Wang Q. Automatic Urinary Sediments Visible Component Detection Based on Improved YOLO Algorithm. 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL) [Internet]. 2020 [cited 2025 Feb 2]. page 485–90. Available from: https://ieeexplore.ieee.org/document/9270436?denied=

10. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [Internet]. arXiv; 2018 [cited 2025 Feb 2]. Available from: http://arxiv.org/abs/1804.02767

11. Chen Z, Hu R, Chen F, Fan H, Ching FY, Li Z, et al. An Efficient Particle YOLO Detector for Urine Sediment Detection. Machine Learning for Cyber Security. Cham: Springer Nature Switzerland; 2023. page 294–308.

12. Li T, Jin D, Du C, Cao X, Chen H, Yan J, et al. The image-based analysis and classification of urine sediments using a LeNet-5 neural network. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. Taylor & Francis; 2020;8:109–14.

13. Pan J, Jiang C, Zhu T. Classification of urine sediment based on convolution neural network. AIP Conference Proceedings. 2018;1955:040176.

14. Khalid ZM, Scddeek Hawezi R, Muhamad Amin SR. Urine Sediment Analysis by Using Convolution Neural Network. 2022 8th International Engineering Conference on Sustainable Technology and Development (IEC) [Internet]. 2022 [cited 2025 Feb 2]. page 173–8. Available from: https://ieeexplore.ieee.org/document/9807482

15. Liu W, Li W, Gong W. Ensemble of fine-tuned convolutional neural networks for urine sediment microscopic image classification. IET Computer Vision. 2020;14:18–25.

16. Nagai T, Onodera O, Okuda S. Deep learning classification of urinary sediment crystals with optimal parameter tuning. Sci Rep. Nature Publishing Group; 2022;12:21178.

17. Ji Q, Li X, Qu Z, Dai C. Research on Urine Sediment Images Recognition Based on Deep Learning. IEEE Access. 2019;7:166711–20.

18. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv [Internet]. 2019;abs/1905.11946. Available from: https://api.semanticscholar.org/CorpusID:167217261

19. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. page 248–55.

20. Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications. 2024;244:122778.

21. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, NSW, Australia: JMLR.org; 2017. page 1321–30.

22. Pakdaman Naeini M, Cooper G, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. Association for the Advancement of Artificial Intelligence (AAAI); 2015;29. Available from: http://dx.doi.org/10.1609/aaai.v29i1.9602

23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017. page 618–26.

24. Maaten L van der, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9:2579–605.

25. PySide6 [Internet]. The Qt Company Ltd; Available from: https://www.qt.io/qt-for-python

26. Gillner S. We’re implementing AI now, so why not ask us what to do? – How AI providers perceive and navigate the spread of diagnostic AI in complex healthcare systems. Social Science & Medicine. 2024;340:116442.

27. Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N. The class imbalance problem in deep learning. Mach Learn. 2024;113:4845–901.

28. Rahmani K, Thapa R, Tsou P, Casie Chetty S, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Med Inform. 2023;173:104930.

29. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognition. 2012;45:521–30.

30. Gama J, Medas P, Castillo G, Rodrigues P. Learning with Drift Detection. In: Bazzan ALC, Labidi S, editors. Advances in Artificial Intelligence – SBIA 2004. Berlin, Heidelberg: Springer; 2004. page 286–95.

31. Deltadahl S, Gilbey J, Laer CV, Boeckx N, Leers M, Freeman T, et al. Deep Generative Classification of Blood Cell Morphology [Internet]. arXiv; 2024 [cited 2025 Apr 14]. Available from: http://arxiv.org/abs/2408.08982

32. Heremans ERM, Seedat N, Buyse B, Testelmans D, van der Schaar M, De Vos M. U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging. Computers in Biology and Medicine. 2024;171:108205.

33. De Lange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis A, et al. A Continual Learning Survey: Defying Forgetting in Classification Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44:3366–85.

34. Goetz L, Seedat N, Vandersluis R, van der Schaar M. Generalization—a key challenge for responsible AI in patient-facing clinical applications. npj Digit Med. Nature Publishing Group; 2024;7:1–4.

TABLES

Table 1
The total collected data including the number of images per class. Dataset A refers to the first collection of data, while the latter batch of more challenging examples is referred to as Dataset B.

Component	Dataset A		Dataset B	
	Count	Percentage	Count	Percentage
Amorphous precipitate	849	2.1%	289	3.4%
Bacteria	5009	12.3%	821	9.7%
Crystals	5195	12.7%	1357	16.0%
Hyaline casts	4553	11.2%	641	7.6%
Mucus	269	0.7%	247	2.9%
Non-squamous epith. Cells (NSEC)	3020	7.4%	609	7.2%
Pathological casts	2974	7.3%	381	4.5%
Red Blood Cells (RBC)	5008	12.3%	931	11.0%
Sperm	160	0.4%	137	1.6%
Squamous epithelium	5000	12.3%	858	10.1%
White blood cells (WBC)	5000	12.3%	1110	13.1%
White blood cell clumps	453	1.1%	116	1.4%
Yeasts	3304	8.1%	966	11.4%
Total	**40794**	100%	**8463**	100%

Table 2
The results in the test data of each experimental scenario. The macro averages of precision, recall and F-1 score are reported. For the third scenario, precision, recall and F-1 Score are reported per class, along with the number of images (support).

Scenario	Precision	Recall	F-1 Score	Accuracy
Scenario 1	0.91	0.96	0.93	0.96
Scenario 2	0.91	0.88	0.90	0.93
Scenario 3	0.95	0.96	0.95	0.97
Scenario 3: Class Specific Results				
Class				**Support (%)**
Amorphous precipitate	0.93	0.95	0.94	114 (2.31%)
Bacteria	0.99	0.98	0.99	583 (11.83%)
Crystals	0.98	0.97	0.98	656 (13.31%)
Hyaline casts	0.97	0.96	0.97	520 (10.55%)
Mucus	0.79	0.92	0.85	52 (1.05%)
Non-squamous epithelial cells	0.92	0.93	0.85	363 (7.36%)
Pathological casts	0.94	0.92	0.93	336 (6.82%)
Red blood cells	0.97	0.97	0.97	595 (12.07%)
Sperm	0.94	1.00	0.97	30 (0.61%)
Squamous epithelium cells	0.98	0.98	0.98	586 (11.89%)
White blood cells	0.96	0.97	0.97	611 (12.39%)
White blood cell clumps	0.91	0.89	0.90	57 (1.16%)
Yeasts	0.99	0.99	0.99	427 (8.66%)

FIGURES

Figure 1. An overview of the study.

Two data sets with different characteristics were collected to create the experimental scenarios. An EfficientNet-based convolutional neural network was developed to evaluate the performance on these scenarios. The best-performing model was integrated into a GUI application deployed in the lab to perform a prospective evaluation.

Figure 2. Reliability diagram.

Model accuracy and average confidence are shown across confidence bins (horizontal axis), with bar height indicating the number of predictions per bin. The Expected Calibration Error (ECE = 0.01) is low, with most predictions being above 0.9 confidence and largely correct, whereas lower-confidence bins show less alignment between accuracy and confidence.

Figure 3. Grad-CAM visualizations of different model predictions.

Each row shows the input image with ground-truth label, top-3 predicted classes with output probabilities, and Grad-CAM heatmaps for the predicted and true class. Arrows in the first row illustrate the derivation of Grad-CAMs; subsequent rows follow the same layout. Correct examples (bottom rows) show only the Grad-CAM of the predicted/true label.

Figure 4. t-SNE projection of model embeddings

The result of projecting the high dimensional test features extracted from the model, in the combined scenario 3, to two dimensions using t-SNE. Points are color-coded according to their predefined classes

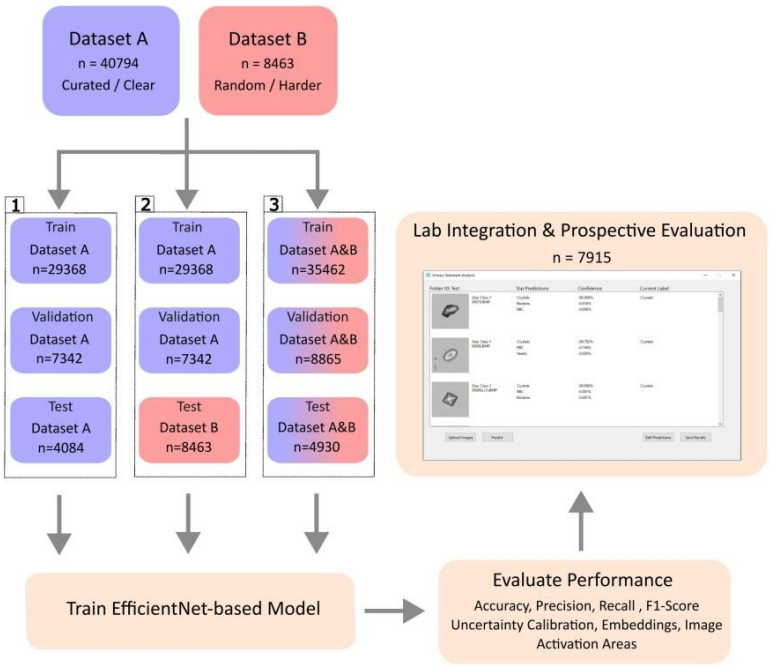


Figure 1. An overview of the study.
Two data sets with different characteristics were collected to create the experimental scenarios. An EfficientNet-based convolutional neural network was developed to evaluate the performance on these scenarios. The best-performing model was integrated into a GUI application deployed in the lab to perform a prospective evaluation.

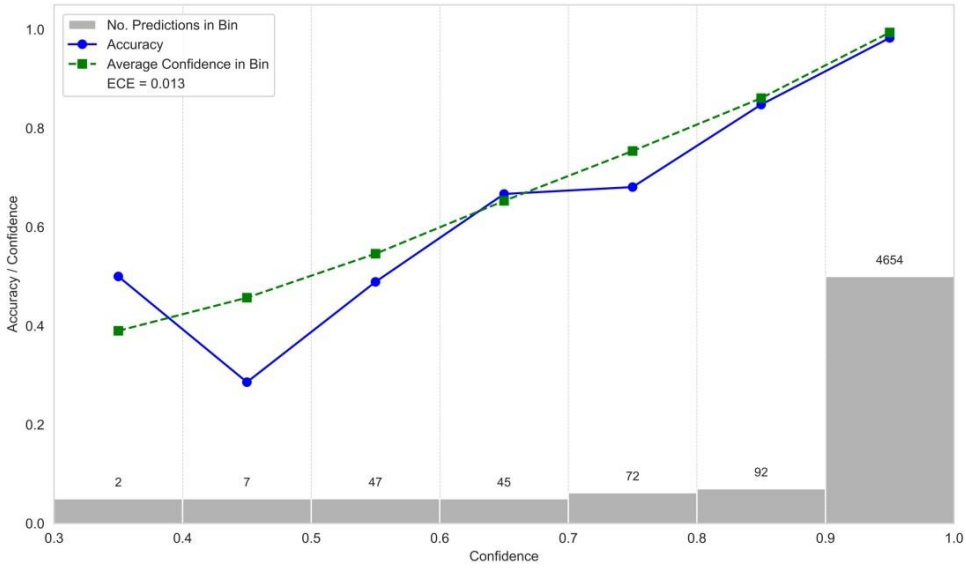


Figure 2. Reliability diagram.
Model accuracy and average confidence are shown across confidence bins (horizontal axis), with bar height indicating the number of predictions per bin. The Expected Calibration Error (ECE = 0.01) is low, with most predictions being above 0.9 confidence and largely correct, whereas lower-confidence bins show less alignment between accuracy and confidence.

259x149mm (300 x 300 DPI)

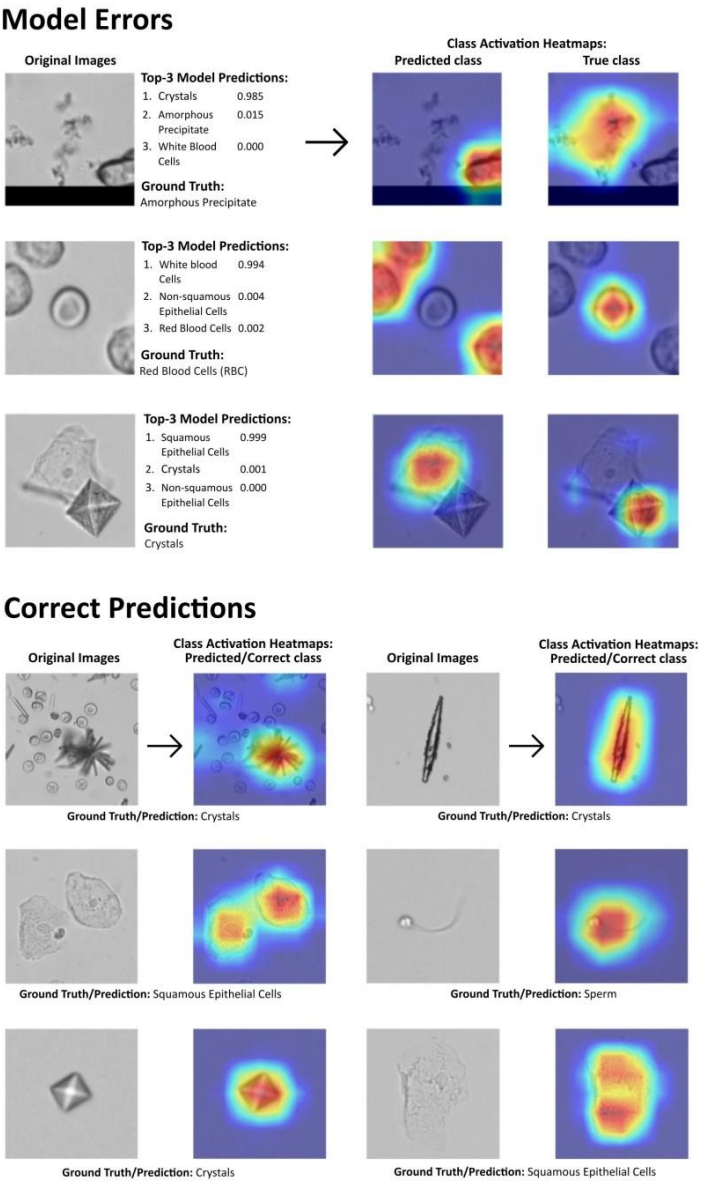


Figure 3. Grad-CAM visualizations of different model predictions.
Each row shows the input image with ground-truth label, top-3 predicted classes with output probabilities, and Grad-CAM heatmaps for the predicted and true class. Arrows in the first row illustrate the derivation of Grad-CAMs; subsequent rows follow the same layout. Correct examples (bottom rows) show only the Grad-CAM of the predicted/true label

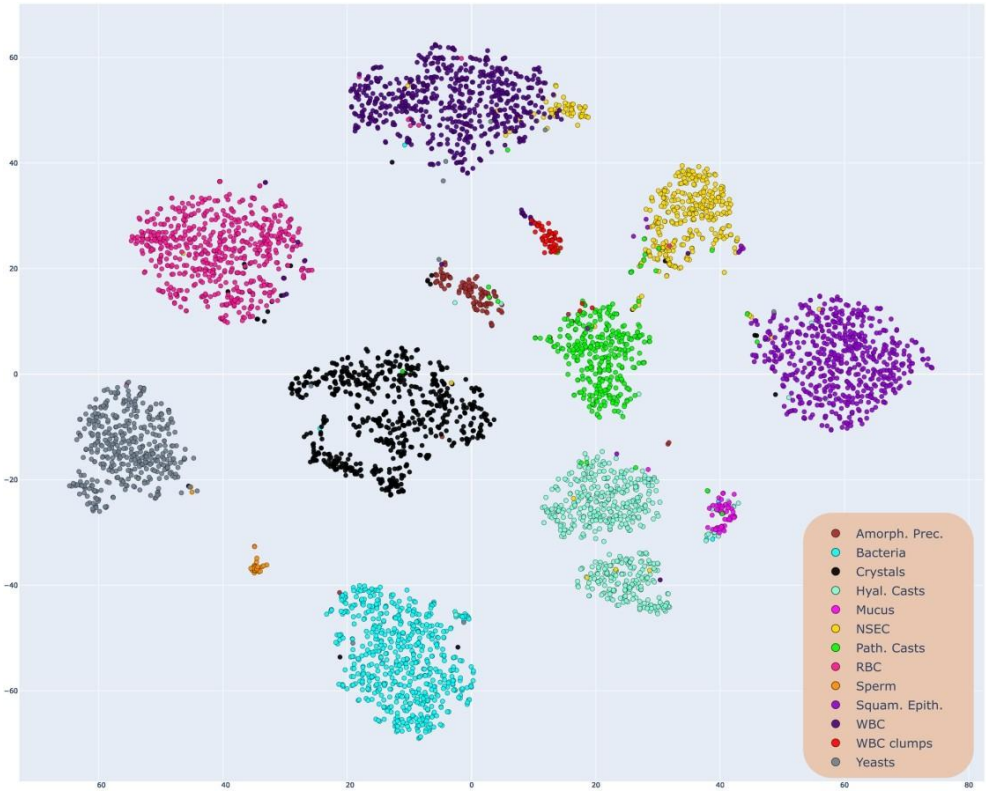


Figure 4. t-SNE projection of model embeddings
The result of projecting the high dimensional test features extracted from the model, in the combined scenario 3, to two dimensions using t-SNE. Points are colour-coded according to their predefined classes