International Journal of Numerical Methods
for Calculation and Design in Engineering

ARTICLE

AI-IoMT Synergy: A Real-Time Framework for Automated Urinary Tract Infections (UTI) Detection Based on Urine Sediments

Abdullahi Umar Ibrahim[1,*], Mohamed Ahmed Mohamed Ahmed[1], Ibrahim Ahmed Ame[2], Chidi Wilson Nwekwo[1], Fathy A A Hassan[1], Suleyman Asir[1], Samuel Nii Tackie[3], Fadi Al-Turjman[2] and John Bush Idoko[4]

[1]Department of Biomedical Engineering, Near East University, Nicosia, Mersin 10, Turkey
[2]AI, Data Analytics, and Software Engineering Departments, AI and IoT research center, AI and Informatics Faculty, Near East University, Nicosia, Mersin 10, Turkey
[3]Department of Electrical and Electronic Engineering, Near East University, North Cyprus, Nicosia, Mersin 10, Turkey
[4]Departments of Computer Engineering/Cyber Security Engineering, Near East University, North Cyprus, Nicosia, Mersin 10, Turkey
*Corresponding Author: Abdullahi Umar Ibrahim. Email: Abdullahi.umaribrahim@neu.edu.tr

Received: 29 December 2025; Accepted: 20 March 2026

ABSTRACT: A Urinary Tract Infection (UTI) is characterized by an infection affecting the urinary system, including the kidneys, bladder, urethra, and ureters, with clinical presentations including pyelonephritis, cystitis, and urethritis. While conventional diagnostic methods such as urinalysis and urine culture and sensitivity (C&S) remain widely used, they are limited by subjectivity, time-intensive processing, susceptibility to contamination and risks of false-positive or false-negative results. This study proposes a comprehensive deep learning (DL) and Internet of Things (IoMT) framework to automate the real-time detection and classification of UTIs using microscopic urine sediment images. The study employed 2 datasets (A and B). Dataset A, a clinically acquired dataset, comprises of 3345 images (normal, erythrocytes, fungi and pus) and Dataset B, a publicly accessible data comprises of 5377 images and 26,419 cropped microscopic images corresponding to 7 classes (casts, crystals, erythrocytes leukocytes, epithelial cells, epithelial nuclei, mycetes). A two-stage classification approach was implemented: a binary task to distinguish urine sediments from normal cases, followed by a multiclass task (clinical data and online data) to identify the specific infection type. All images underwent pre-processing, including normalization, resizing, noise removal, and augmentation to enhance feature visibility and model generalizability. The data were partitioned into training (65%), validation (25%), and test (10%) sets. Six state-of-the-art DL architectures, including ResNet50-V2, ResNet101-V2, Inception-V3, Xception-V2, Inception-ResNet-V2, and VGG19 were fine-tuned using transfer learning and evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The proposed models were uploaded to a website to enable real-time detection (accessible via this link: https://uticlassification.app/). The proposed pipeline demonstrated strong performance in both classification tasks, underscoring the potential of deep learning as a reliable, rapid, and reproducible tool for automated urine sediment identification in clinical practice.

KEYWORDS: Urine sediments; urinary tract infection (UTI); artificial intelligence; deep learning; transfer learning; Internet of Medical Things; medical imaging

1 Introduction

A Urinary Tract Infection (UTI) is an infection that affects the urinary tract or any part of the urinary system (e.g., bladder, kidney, urethra, etc.). Even though the symptoms of UTIs may differ based on the primary organ, some of the general signs and symptoms of UTI include a persistent urge to urinate, painful urination (Dysuria), frequent urination of small amounts, foul-smelling or cloudy urine, etc. UTIs represent a significant burden on healthcare systems and are a frequent cause of emergency department visits [1].

It is estimated that 50% of the population will experience a UTI in their lifetime, and they are particularly common in women (i.e., 50–60% of women will have at least one episode) [2]. It was also reported that approximately 404.61 million people were experiencing UTIs in 2019 [3,4]. While another report estimated the number of cases to reach approximately 4.49 billion (including recurrent episodes) [5]. Consequently, it is also estimated that the global lifetime probability of developing UTI or related disease, such as kidney, inflammation, is extremely high (i.e., over 93%, 96% for females and 77% for males) [6].

The detection of urine particles or sediments is essential for diagnosing UTIs and other kidney disorders, since the identification of red blood cells (RBCs), white blood cells (WBCs), casts, fungi, and bacteria can signal conditions like hematuria, kidney stones, and urinary tract cancers [7]. Thus, in order to diagnose a patient suspected of UTIs, medical experts rely on clinical symptoms, urinalysis and confirmatory urine culture. However, each method has its own significant drawbacks. For instance, manual microscopy is time-consuming, technically difficult, and prone to inter-observer error. Moreover, urinalysis and microscopy are not widely available as a point-of-care test. Despite the fact that urine culture is regarded as the gold standard test, it is limited by low sensitivity, time-consuming, susceptibility to contamination, etc. [8–10].

Aside from conventional techniques, clinical pathologists also utilize advanced molecular techniques such as Next-Generation Sequencing (NGS) and Polymerase Chain Reaction (PCR). Although these techniques offer much higher sensitivity and can identify a broader range of pathogens in hours instead of days, they are more expensive and, due to high sensitivity, they can detect bacteria even in healthy individuals (i.e., asymptomatic bacteriuria), potentially leading to poor specificity and overtreatment if not interpreted carefully [11,12].

Given the limitations of current methods, there is a need for faster, more accurate, and accessible diagnostics. One promising solution that has attracted significant interest from various researchers is the application of Artificial Intelligence (AI) and Machine Learning (ML), particularly deep learning (DL), into the diagnostic workflow. AI-based techniques such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), Random Forest (RF), etc., are trained to analyze diverse data points and predict the likelihood of diseases [13,14]. Computer-aided detection (CAD) designed using ML and DL models is trained using urine microscopy images in order to identify and count erythrocytes, leukocytes, epithelial cells, and bacteria. Thus, the integration of CAD in the detection of UTI provides rapid, scalable and fast screening, while eliminating human error [15,16].

Despite the promise of AI in aiding the diagnosis of UTIs based on urine sediment identifications, the majority of existing AI models are restricted to clinical settings or laboratories, therefore, they are not readily accessible or available for users. In order to combat this challenge, there is a need to develop a framework that integrates CAD/AI with Internet of Things (IoT) or Internet of Medical Things (IoMT) for real-time detection. Thus, this study addressed these limitations as follows:

- Proposed a dual-task diagnostic framework that performs both binary classification (urine sediments vs. normal cases) and 2 multiclass classifications (fungi, pus cells, and RBCs) and (casts, crystals, erythrocytes leukocytes, epithelial cells, epithelial nuclei, mycetes).
- Use of a hospital-acquired dataset which comprises 3372 images sourced from a clinical setting, moving beyond reliance on public repositories.
- Use of a public-acquired dataset that comprises 5377 images and 26,419 cropped microscopic images.
- Comprehensive model evaluation of pretrained CNNs, including VGG19, ResNet, Inception, and Xception for both classification tasks using conventional metrics.
- Comprehensive framework evaluation based on numerical stability and computational performance.

2 Literature Review

Over the years, several studies have attempted to deploy AI-based models for the detection and classification of UTIs based on the identification of different types of urine sediments. One such study was proposed by Naznine et al. [7]. The study reported the deployment of a DL-based technique for the detection of UTIs. The overall methodology relied on an ensemble framework that combines YOLOv9e and KD-

YOLOX-ViT (which utilizes Knowledge Distillation and Vision Transformers) using a Weighted Box Fusion (WBF) technique. The proposed technique is trained and tested using the Urine Sediment Dataset (USE), with a total number of 5376 images categorized into seven distinct cell types: casts, crystals, epithelial cells, epithelial nuclei, erythrocytes, leukocytes and mycetes. Performance evaluation of the framework resulted in a mean Average Precision (mAP50) of 94.18% and 94.64% on the external dataset.

Lyu et al. [17] reported the development of YUS-Net, an advanced detection model based on the YOLOX architecture. In order to improve detection accuracy, the model utilizes a specific data augmentation strategy, pre-trained weights, attention modules, and a novel "Varifocal" loss function. The proposed framework is trained and tested using the USE dataset, which consists of 5376 filtered microscopic urine images distributed across 7 classes. Assessment of the performance of the proposed YUS-Net resulted in an accuracy of 96.07%, mAP of 96.07% and fast processing speed of 26.13 ms per image.

The study reported by Wang et al. [18] deployed a DP-YOLO, a YOLOv5s enhanced with two modifications, including Petal-like Sample Amplification (PSA) and DYB, a backbone incorporating deformable convolutions (DCNv2 & DCNv3). The proposed enhanced DP-YOLO was evaluated on the COCO2017 dataset and the domain-specific Urised11 urine sediment dataset (7364 images, 58,196 instances, 11 classes). Performance evaluation of the proposed system resulted in 41.2% AP (a +3.2% improvement over YOLOv5s) on the COCO dataset and 49.2% AP (76.2% AP_{50}) on the Urised11 dataset.

Due to the time-consuming and subjectivity nature of manual microscopy, Liou et al. [19] proposed a novel DL model known as "Patch U-Net," designed to recognize specific cell types. The authors also created and trained the model using an open urine microscopy dataset containing 300 images with 3562 manually annotated cells across seven classes (e.g., rods, RBC/WBC). Performance evaluation of the proposed Patch U-Net on the test set resulted in a dice coefficient of 0.877, precision of 0.890, recall of 0.920, and an AUC of 0.989.

Considering the fact that DL models such as CNNs, typically require a large amount of manually annotated data, leading to high computational complexity, Suhail and Brindha [20] proposed the implementation of five variants of You Only Look Once version 5 (YOLOv5) (YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) for the detection of six urine sediments (erythrocyte, leukocyte, crystals, cast, mycetes and epithelial cells) from 5376 microscopic images. Moreover, model training was enhanced using an Evolutionary Genetic Algorithm (EGA) for hyperparameter optimization. Comparative analysis of the performance of 5 YOLOv5 variants indicated that YOLOv5l achieved the best result with 85.8% mAP @ IoU 0.5 and 23.4 ms detection speed per image.

Liang et al. [21] proposed a model known as Feature Pyramid Network with DenseNet (DFPN) for automated urine sediment examination. The proposed model is trained on the USE dataset, containing 5377 images with 42,759 labeled instances across 7 cell types. Evaluation of the proposed DFPN on the USE test set resulted in a final mAP of 86.9%. In order to address the challenge of low-resolution images and improve recognition accuracy, Avci et al. [22] proposed a new super-resolution Faster Region-based CNN (Faster R-CNN) method based on three different backbones: AlexNet, VGG16, and VGG19. These backbone hybrid models are trained using 500 color urine images (51,077 annotated patches) across 10 different urine sediments. Comparative assessment of the proposed methods indicated that AlexNet-based Faster R-CNN achieved the best result with 98.6% classification accuracy.

Li et al. [23] proposed a DL-based approach using RetinaNet with a ResNet50 backbone and FPN for the detection of urine sediment. The proposed framework was trained and tested using a custom dataset of 15,360 urine images labeled across seven cellular components (e.g., erythrocytes, leukocytes, epithelial cells, epithelium, casts, crystal and squamous epithelial cells). Assessment of the performance of the framework resulted in an accuracy of 88.65% and a processing time (i.e., inference speed) of 0.2 s per image. The overall summary of the existing study is presented in Table 1.

Table 1: Summary of Existing Studies.

Ref.	Dataset + Size	# Classes	Models	Result
[7]	USE (5376 images)	7	Ensemble (YOLOv9e and KD-YOLOX-ViT)	94.18% mAP
[17]	USE (5376 images)	7	YUS-Net	96.07% accuracy and 96.07 mAP
[18]	Urised11 (7364 images)	11	DP-YOLO	41.2% AP on COCO2017 and 49.2% AP on Urised11
[19]	300 images (3562 annotated cells)	7	Patch U-Net	Precision of 0.890, Recall of 0.920, and an AUC of 0.989
[20]	5376 images	6	YOLOv5	85.8% mAP @ IoU 0.5
[21]	USE dataset (5377 images)	7	DFPN	mAP of 86.9%.
[22]	500 images (51,077 annotated patches)	10	AlexNet-based Faster R-CNN	98.6% accuracy
[23]	15,360 images	7	RetinaNet	88.65% accuracy

Note: mAP: Minimum Average Precision.

Limitations of Existing Studies

Automated urine microscopy is crucial for diagnosing kidney and urinary issues, yet standard methods often struggle to identify sediment particles due to blurry boundaries, sample imbalances, and visual variations. Several studies have attempted to detect urine sediments using DL-based methods. Despite the growing number of these studies, several limitations persist. Among these limitations is the use of common datasets or overlapping datasets (e.g., USE dataset with 5376 images, Urised11 with 7364 images). Moreover, the majority of existing studies rely on DL models (such as YOLO variants, Faster R-CNN and U-Net) for urine sediment classification/detection. Additionally, many studies report standard metrics like mAP and accuracy.

In order to address these limitations, this study proposed a dual-task framework that performs both binary classification (urine sediments vs. normal cases) and 2 multiclass classification tasks (specific fungi, pus, RBCs) and (casts, crystals, erythrocytes leukocytes, epithelial cells, epithelial nuclei, mycetes). Moreover, the study also utilized a hospital-acquired dataset (3372 images), not just public repositories. Consequently, this study also deployed multiple CNNs (VGG19, ResNet50, ResNet101, Inception, InceptionResNet and Xception) for both binary and multiclass tasks. Lastly, this study developed an AI/IoT-based framework for real-time identification of clinically important urine sediments. The comparison between the limitations of existing studies and the research gap explored in this study is presented in Table 2.

Table 2: Comparison between the limitations of existing studies and the research gap.

Limitations of Existing Studies	Research Gap Explored in This Study
Most studies only identified different types of urine sediments	Explicitly builds a diagnostic pipeline: first detects discriminates between urine sediments and normal cases (binary), then classifies type (multiclass)
Limited to public datasets	Uses a hospital-sourced dataset and publicly-sourced dataset
Few studies compare multiple modern architectures	Compares 6 CNNs (VGG19, Xception, ResNet50, ResNet101, Inception and InceptionResNet)
Existing studies proposed offline framework (i.e., restricted to clinical settings or research)	Developed website that support real-time identification of urine sediments

Relies on conventional evaluation metrics	Include other evalaution techniques such as numerical analysis and computational performance test

3 Experimental Set-Up

This study proposed an AI/IoT-based framework for the binary and two multiclass classifications of urine sediments. The study is designed based on a pipeline that includes curation of clinical dataset, image pre-processing, data split, data augmentation, training and testing of pre-trained CNNs including InceptionResNetV2, InceptionV3, ResNet50, ResNet101, VGG19 and Xception, performance evaluation, development of a website and deployment of models for real-time detection. The overall summary of the methodology is presented in Fig. 1.

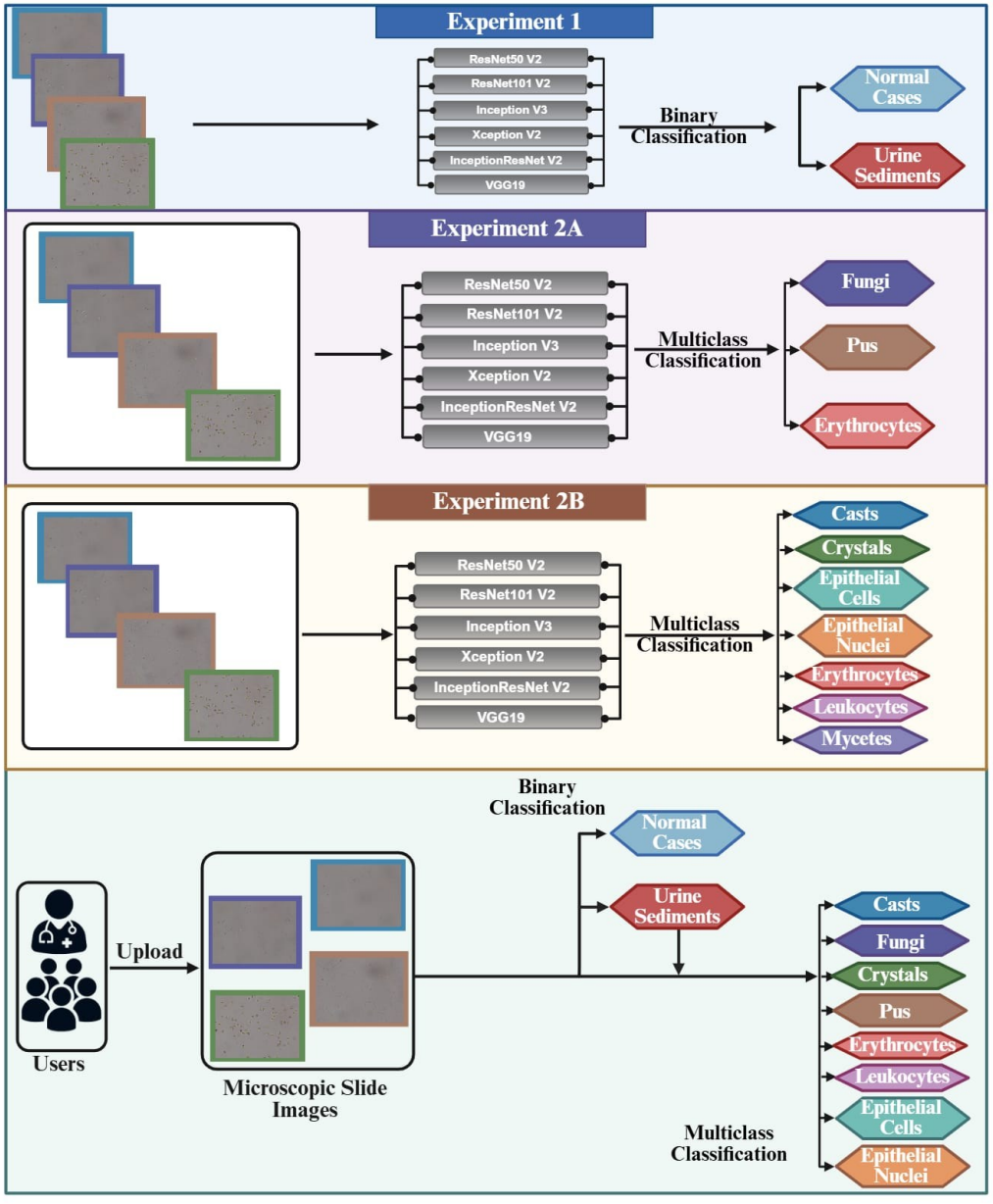
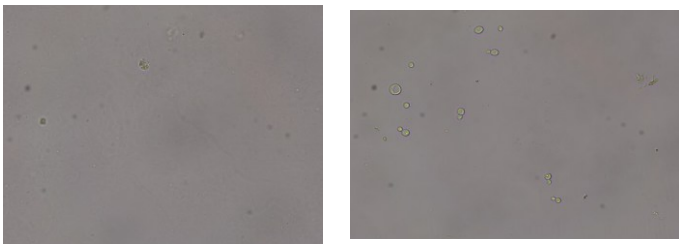


Figure 1: Proposed Pipeline.

3.1 Data Collection

3.1.1 Binary

The binary dataset was constructed with two distinct classes: urine sediment and no urine sediment, as shown in Fig. 2. The no urine sediment class comprises 498 images after duplicate removal. The urine sediment class comprises 1000 images, including 300 images from a clinically-acquired dataset (100 from each class: erythrocytes, pus cells and fungi) and 700 images from the Urine Sediment Dataset (USE) distributed across seven classes: cast, crystal, epithelial cells, epithelial nuclei, erythrocytes, leukocytes and mycetes (i.e., 100 from each class). In order to balance the no urine sediment, offline data augmentation was conducted via horizontal flips and 180-degree rotations applied randomly with a probability of 0.5, thereby increasing the dataset to 1000 images.



(**a**) (**b**)

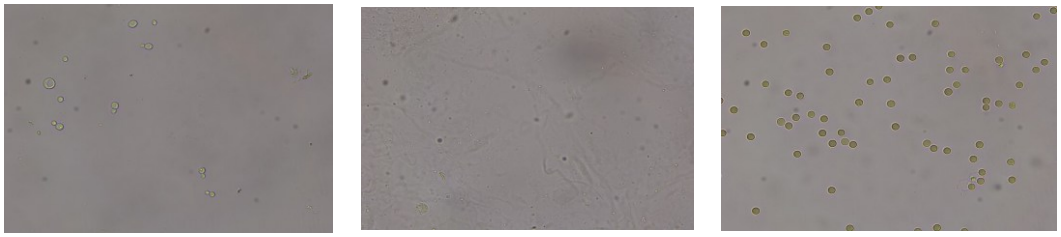
Figure 2: Samples (**a**) Urine sediment (**b**) Normal.

3.1.2 Multiclass

Dataset A

The clinically acquired dataset (i.e., Dataset A) was obtained from the BEN RUCHD Laboratory, Tobruk, Libya, between January 2019 and December 2024. The dataset comprises microscopic images of urine sediment samples collected from patients presenting with symptoms suggestive of urinary tract infection (such as dysuria, frequency, urgency and suprapubic pain). The cohort consisted of 847 patients (582 female, 265 male). Patient ages ranged from 18 to 87 years (mean age: 47.3 ± 16.8 years). The inclusion criteria include: adults (i.e., ≥18 years) patients with clinical suspicion of UTI, availability of a fresh urine sample for microscopy and provision of informed consent. The only exclusion criteria include pregnant women, menstruating women and patients on antibiotic therapy within the preceding 48 h. 3 independent laboratory technologists with more than 5 years of experience in urinalysis performed manual annotation of all images. Each image was labeled according to the predominant sediment type. Erythrocytes (characterized by biconcave and discoid morphology), pus cells (neutrophils or granular cytoplasm with multilobed nuclei) and fungi (i.e., Candida species, characterized by oval budding yeast cells). Disagreements between laboratory technologies were resolved by consensus review with a senior pathologist. Inter-annotator agreement, assessed using Fleiss' kappa on a subset of 300 images, was 0.89 (95% CI: 0.84–0.94), indicating excellent agreement.

Following the removal of duplicate images, the final dataset comprises 2847 total images, 949 images per category. This configuration was designed to train models capable of discriminating between 3 specific urine sediments based on microscopic features. The aim was to enable a more refined diagnostic approach by sub-classifying specific urine sediments, thereby supporting the identification of the infection's nature and probable source. The samples of the multiclass dataset are presented in Fig. 3.



(**a**) (**b**) (**c**)

Figure 3: Samples of Multiclass Dataset A: (**a**) Fungi (**b**) Pus and (**c**) Erythrocytes.

Dataset B

The study also acquired a publicly accessible dataset known as the Urine Sediment Dataset (USE). The dataset consisted of 5376 annotated images corresponding to 7 categories of urinary sediment particles. The dataset is divided into 4256 images for training, 852 images for validation, and 268 images for testing. Moreover, the dataset contains 26,419 cropped microscopic images grouped into seven classes: Cast (3662), Crystal (1644), Epithelial cell (6175), Epithelial nuclei (667), erythrocytes (21,815), leukocytes (6169), and Mycetes (2083). The erythrocytes class was downsampled to 6000 images to reduce class imbalance and improve training stability. The dataset can be accessed via: https://github.com/174614361/Urinary-Sediment-Dataset?tab=readme-ov-file. Each sample of the seven classes is presented in Fig. 4.

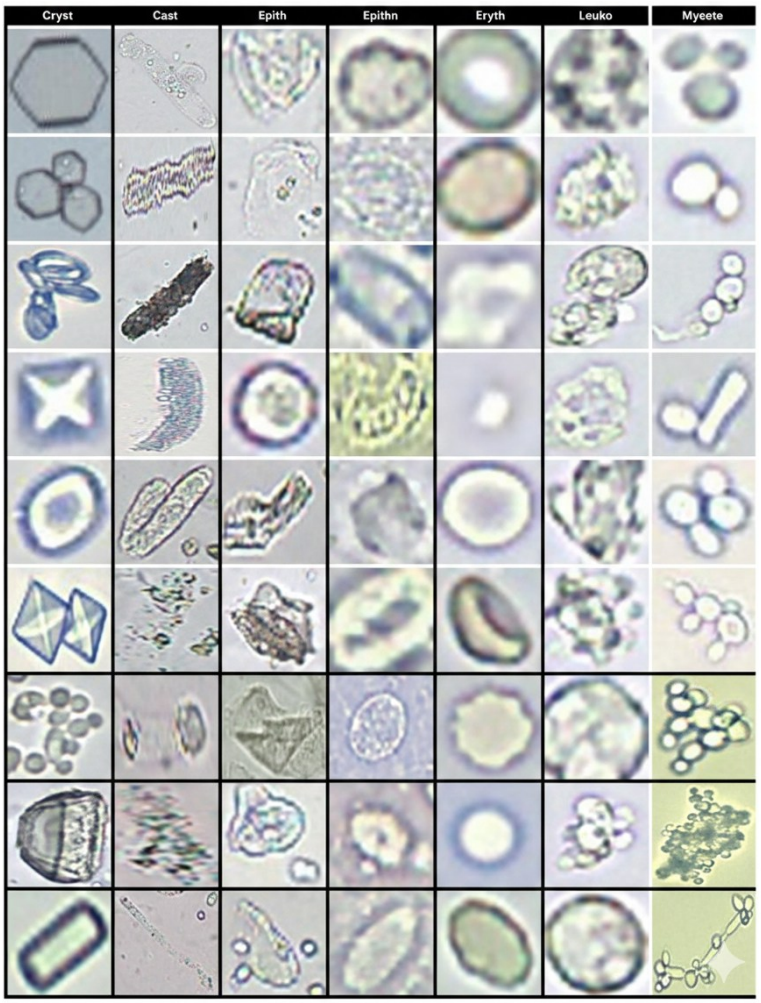


Figure 4: Samples of Multiclass Dataset B.

3.2 Data Split

Data splitting is critical for ensuring robust AI model development. The two most common data splits include 2 partitions (training and testing or training and validation and 3 partitions (training, validation, and testing sets). The training set is used to learn parameters, the validation set tunes hyperparameters and prevents overfitting, while the testing set provides an unbiased evaluation of final model performance. This separation is crucial for assessing generalizability, preventing data leakage, and ensuring the resulting accuracy reflects real-world predictive capability on unseen data. Thus, in this study, the acquired dataset for both binary and multiclass classifications was divided into three, namely 65% training, 25% validation, and 10% testing, respectively as summarized in Table 3.

Table 3: Data split for Binary Class.

Binary Classification				
Class	**Training (65%)**	**Validation (25%)**	**Testing (10%)**	**Total Images**
No Urine Sediment	324	125	49	498
Urine Sediment	324	125	49	498
Total	**648**	**250**	**98**	**996**
Multiclass: Dataset A				
Class	**Training (65%)**	**Validation (25%)**	**Testing (10%)**	**Total Images**
Fungi	617	237	95	949
Pus	617	237	95	949
RBCs	617	237	95	949
Total	**1851**	**711**	**285**	**2847**
Multiclass: Dataset B				
Class	**Training (65%)**	**Validation (25%)**	**Testing (10%)**	**Total Images**
Casts	2376	912	374	3662
Crystals	1069	411	164	1644
Epithelial Cells	4014	1544	617	6175
Epithelial Nuclei	433	167	67	667
Erythrocytes	3912	1505	602	6000
Leukocytes	4010	1542	617	6169
Mycetes	1354	5,21	208	2083
Total	**17,166**	**6602**	**2649**	**26,419**

3.3 Image Pre-Processing

Image pre-processing is essential for addressing common issues in medical imaging datasets, such as noise, artifacts, overlap, low resolution and blur. Addressing these issues can enhance the reliability and effectiveness of AI-based analysis. Thus, in this study, several pre-processing steps were implemented. First, images were reorganized according to their class labels. Each image was subsequently normalized to rescale pixel values from a 0–255 range to 0–1, reducing inter-sample variance. Subsequently, all images were resized to a uniform dimension of 224 × 224 × 3 pixels via resampling to meet model input requirements. Image enhancement techniques were also applied to improve clarity and feature visibility. Lastly, the processed images were converted into NumPy arrays, with pixel data assigned to the feature matrix (x) and corresponding class labels to the target vector (y), preparing the dataset for model training.

3.4 Data Augmentation

During training, data augmentation techniques were applied to improve generalization. These included random flips, rotation at a 0.1 ratio, zoom at a 0.1 ratio and contrast adjustment at a factor of 0.2. Augmentation was performed dynamically during training over 35 epochs, ensuring that the model encountered slightly varied versions of the images in each epoch.

3.5 Pre-Trained Models

3.5.1 VGG

VGG was introduced in 2014 by two scientists (Karen Simonyan and Andrew Zisserman) from the University of Oxford. The key architectural principle of the network revolves around the replacement of large convolutional filters (like the 11 × 11 and 5 × 5 in AlexNet) with stacks of multiple 3 × 3 filters. VGG-16 was ranked second in the ILSVRC 2014 with 7.0% top-5 error rate, trailing behind GoogleNet with 6.67% error. The two variants of VGG include VGG-16 and VGG-19. In terms of architectures, both share the same design philosophy of using only 3 × 3 convolutions and 2 × 2 max-pooling. The main difference is depth, where VGG19 is extended by 3 convolutional layers. The development of VGG pioneered the use of very deep, simple convolutional stacks and also served as a foundation for upcoming networks.

3.5.2 ResNet

The Residual Learning Framework or Residual Network (i.e., ResNet) is a neural network developed in order to address the issue of degradation. This problem arises as networks get deeper, accuracy saturates and then degrades rapidly due to vanishing/exploding gradients. In order to address this issue, scientists (Kaiming He and colleagues) introduced Residual blocks with skip connections (shortcuts) to enable Identity mapping and uninterrupted gradient flow through shortcuts. ResNets were ranked in 1st place in the ILSVRC 2015 classification task with a top-5 error rate of 3.57%. The original ResNet Architectures include ResNet18 and 34. Subsequently, the depth was increased to 50, 101 and 152. Other key variants include ResNet-v2, Wide ResNet, ResNeXT, ResNet-D, etc. The basic building block, ResNet, is the "bottleneck" design. The model stacks three convolutional layers, first a 1 × 1 convolution to reduce dimensionality, followed by a 3 × 3 convolution for processing, and finally another 1 × 1 convolution to restore dimensionality. In this study, two variants of ResNet are used, including ResNet50 and 101. The rationale behind the selection of these two variants is due to their low computational parameters (25.6M and 44.5M) compared with ResNet152 (60.2M).

3.5.3 Inception ResNet

Inception ResNet is a hybrid architecture that integrates the strengths of Inception modules (for multi-scale feature extraction) and ResNet's skip connections (for stable training of very deep networks). The key innovation of the network revolves around the addition of a residual (skip) connection to each inception module. A 1 × 1 convolution scales up filter bank depth after each module. The model architecture is structured based on streamlined multiple Inception-ResNet blocks and reduction modules. Some of the variants of the network include Inception-ResNet v1 and Inception-ResNet v2.

3.5.4 Inception

Unlike ResNet, which focuses on addressing the issue of vanishing gradient in very deep networks, the Inception network's key innovation focuses on solving the issue regarding traditional CNNs, where choosing a single convolution filter size (such as 3 × 3 vs. 5 × 5) forces a trade-off. To address this issue, the Inception module performs multiple convolutions with different filter sizes (such as a 1 × 1, a 3 × 3 and a 5 × 5) in parallel on the same input and concatenates their outputs. This enables the network to capture multi-scale features within a single layer. Moreover, to reduce dimensionality, 1 × 1 convolution is added before 3 × 3 and 5 × 5. The inception network was introduced by scientists working at Google (i.e., Szegedy and colleagues). Inception v1 was ranked first in the ImageNet Large Scale Visual Recognition Challenge in 2014 with a top-5 error rate of 6.67%. In terms of architectures, the model comprises 22 layers (9 Inception modules). Inception has several variants, including inception v1, v2, v3, v4 and Inception-ResNet.

3.5.5 Xception

Xception or Extreme Inception, introduced in 2017 by François Chollet, is a deep convolutional neural network designed to compete with Inception v3 due to their shared similarities in terms of network parameters (around 22–23 million), while ensuring any performance difference is due to architectural

efficiency. The key innovation of Xception revolves around the replacement of the complex multi-branch Inception modules with streamlined depthwise separable convolution layers, leading to a simpler, efficient and powerful architecture. The architecture of the network consists of 36 convolutional layers structured into three sequential flows; entry, middle and exit flows.

3.6 Training and Fine-Tuning Parameters

In order to train the proposed models, the experimental setup involved configuring the computational environment, defining hyperparameters, loading the dataset, and initiating the training process. This procedure was conducted in two stages. In the first stage, an initial verification phase was performed offline on a local workstation to ensure that the Jupyter notebooks executed without errors and generated the correct model files. For efficiency during this verification, models were run for only one epoch. After successful validation, the notebooks and data were compressed and transferred to a cloud-based server for full-scale training.

Training was conducted on servers provisioned through vast.ai, a platform for outsourced high-performance computing. A pre-configured TensorFlow template was used to instantiate the required environment, which included all necessary software packages, a Jupyter notebook interface and tools for data upload and model retrieval. The server specifications included 48 GB of RAM, an RTX 4050 GPU, and 16 GB of SSD storage. To expedite experimentation, three identical servers were employed in parallel, each hosting the complete codebase and dataset for all model variations.

3.6.1 Transfer Learning Configuration

The experimental design encompassed six distinct model architectures, including ResNet50 V2, ResNet101 V2, Inception V3, Xception V2, Inception-ResNet V2, and VGG19, each trained on three separate datasets (binary and two multiclass tasks). To account for variability, each of the eighteen architecture–dataset combinations was trained five times with different random initializations, resulting in a total of ninety trained models. This approach provided varied starting points for each run, thereby producing a range of metric outcomes for more robust evaluation. To ensure full reproducibility, all training parameters and fine-tuning strategies are explicitly detailed below:

All pre-trained models (ResNet50-V2, ResNet101-V2, Inception-V3, Xception-V2, Inception-ResNet-V2, and VGG19) were initialized with ImageNet weights. The top classification layers were replaced with custom layers adapted to each classification task:

Binary: Flatten → Dense (2, sigmoid)

Multiclass Dataset A (3 classes): GlobalAveragePooling2D → BatchNormalization → Dense(512, ‘ReLU’) → Dropout(0.5) → Dense(256, ‘ReLU’) → Dropout(0.3) → Dense(3, ‘softmax’)

Multiclass Dataset B (7 classes): GlobalAveragePooling2D → BatchNormalization → Dense(512, ‘ReLU’) → Dropout(0.5) → Dense(256, ‘ReLU’) → Dropout(0.3) → Dense(7, ‘softmax’)

3.6.2 Training Hyperparameters

Hyperparameters were adjusted dynamically during training based on interim results. If model performance was unsatisfactory upon completion of a run, the scripts were modified, and the training was repeated. After all training concluded, all code, data, and model files were synchronized to Google Drive to maintain backup copies and ensure data redundancy. Finally, performance metrics, including accuracy, loss, precision (Pc), recall (Rc), F1-score, specificity (Sp), Cohen’s Kappa (CK), Matthews Correlation Coefficient (MCC) and AUC were computed across all runs. The top-performing models for each architecture and task were selected for subsequent integration into the web application for prediction. All models were trained using the following hyperparameters, as shown in Table 4, optimized through grid search on validation data.

Table 4: Training Hyperparameters.

Parameter	Binary	Multiclass Dataset A	Multiclass Dataset B

Optimizer	Adam	Adam	Adam
Learning rate	0.000001	0.000001	0.000001
Batch size	32	32	32
Epochs	35	35	35
Loss function	Binary cross entropy	Categorical cross entropy	Categorical cross entropy

3.6.3 Layer Freezing Strategy

A one-stage fine-tuning approach was employed. All base model layers were frozen, and only the newly added classification heads were trained for 35 epochs. This approach ensured that the learned features of the base models were preserved while also preventing overfitting to the dataset.

3.6.4 Data Augmentation Parameters

Augmentation was applied dynamically during training using TensorFlow's ImageDataGenerator with the following parameters:

python

tf.keras.layers.RandomFlip("horizontal_and_vertical"),

tf.keras.layers.RandomContrast(factor = 0.2),

tf.keras.layers.RandomZoom(0.1),

tf.keras.layers.RandomRotation(0.1)

3.6.5 Hardware and Software Environment

Hardware: NVIDIA RTX 4050 GPU (16 GB VRAM), 48 GB RAM, 16 GB SSD storage

Software: Python 3.9, TensorFlow 2.18.0, Keras 3.8.0, CUDA 12.5, cuDNN 9.3

Cloud platform: Vast.ai instances with identical configurations for parallel training

3.6.6 Random Seed Configuration

For reproducibility, all random operations were seeded during the preparation phase. The seed was then removed to allow ANOVA training. The random seed was set to 42 for that phase.

```
python
import numpy as np
import tensorflow as tf
import random
random.seed(42)
np.random.seed(42)
tf.random.set_seed(42)
```

3.7 Statistical Analysis

A statistical method based on Analysis of Variance ANOVA is used to assess significant differences between the means of models trained after five rounds. The result of the rounds is compiled based on evaluation metrics after starting at a different starting point for each round.

4 Results and Discussion

4.1 Performance Evaluation of Models Deployed for Binary Classification

Binary classification, or accurate discrimination between urine sediment and normal cases, is critical for proper treatment. Thus, in this study, we deployed several pre-trained CNNs, including InceptionResNetV2, InceptionV3, ResNet50, ResNet101, VGG19 and Xception. The models are evaluated using several metrics, including Precision, Recall, F1-Score, Specificity, AUC and Test Accuracy. The performance results of the models are presented in Table 5. The accuracy graph of the 6 deployed models is presented in Fig. 5. The ROC graphs are presented in Fig. 6. The results of all deployed models based on 5 rounds are presented in the supplementary file "S1".

Table 5: Performance Evaluation of Models Deployed for Binary Classification.

Model Name	Pc	Rc	F1-Score	Sp	AUC	Accuracy	MCC	CK
InceptionResNetV2	98.03	99.4	98.71	98.00	100.00	98.70	97.41	97.40
InceptionV3	98.06	100.00	99.01	98.00	100.00	99.00	98.03	98.00
ResNet50	98.05	99.60	98.81	98.80	100.00	98.88	97.62	97.60
ResNet101	97.28	99.60	98.42	97.20	100.00	98.40	96.83	96.80
VGG19	98.09	99.20	98.62	97.80	100.00	98.60	97.24	97.20
Xception	99.21	99.60	99.20	99.20	100.00	99.40	98.80	98.80

Note: Pc: Precision; Rc: Recall; Sp: Specificity; AUC: Area under the Curve; Matthews Correlation Coefficient (MCC); Cohen's Kappa (CK).

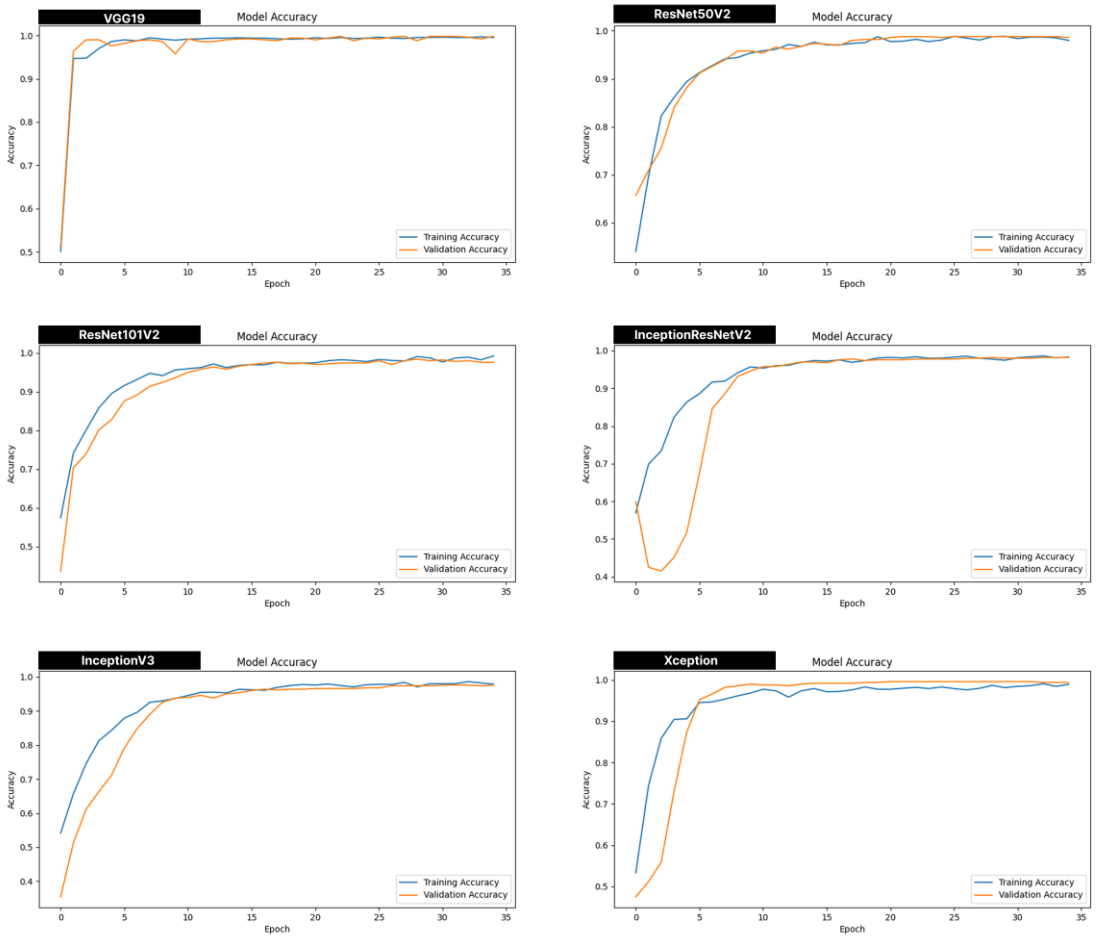


Figure 5: Accuracy graphs of Models deployed for Binary Classification.

Confusion Matrix of Models Deployed for Binary Classification

To test the performance of the deployed models, 200 images (100 normal cases and 100 urine sediments) were selected. The performance evaluation of the models has shown that InceptionResNetV2 accurately identified 198/200 cases, resulting in 2 miss-classifications. InceptionV3 on the other hand, accurately identified 199/200, resulting in 1 miss-classification. Consequently, performance evaluation of ResNet50 indicated that the model accurately identified 199/200 cases, resulting in 1 miss-classification. Subsequently, evaluation of ResNet101 revealed that the model accurately identified 198/200 cases, resulting in 2 miss-classifications. Likewise, VGG19 accurately identified 199/200 cases, resulting in 1 miss-classification. Lastly, evaluation of Xception implied that the model accurately identified all the 200/200 cases, resulting in 0 miss-classification. The Confusion matrix of models deployed for binary classification is summarized in Fig. 7.

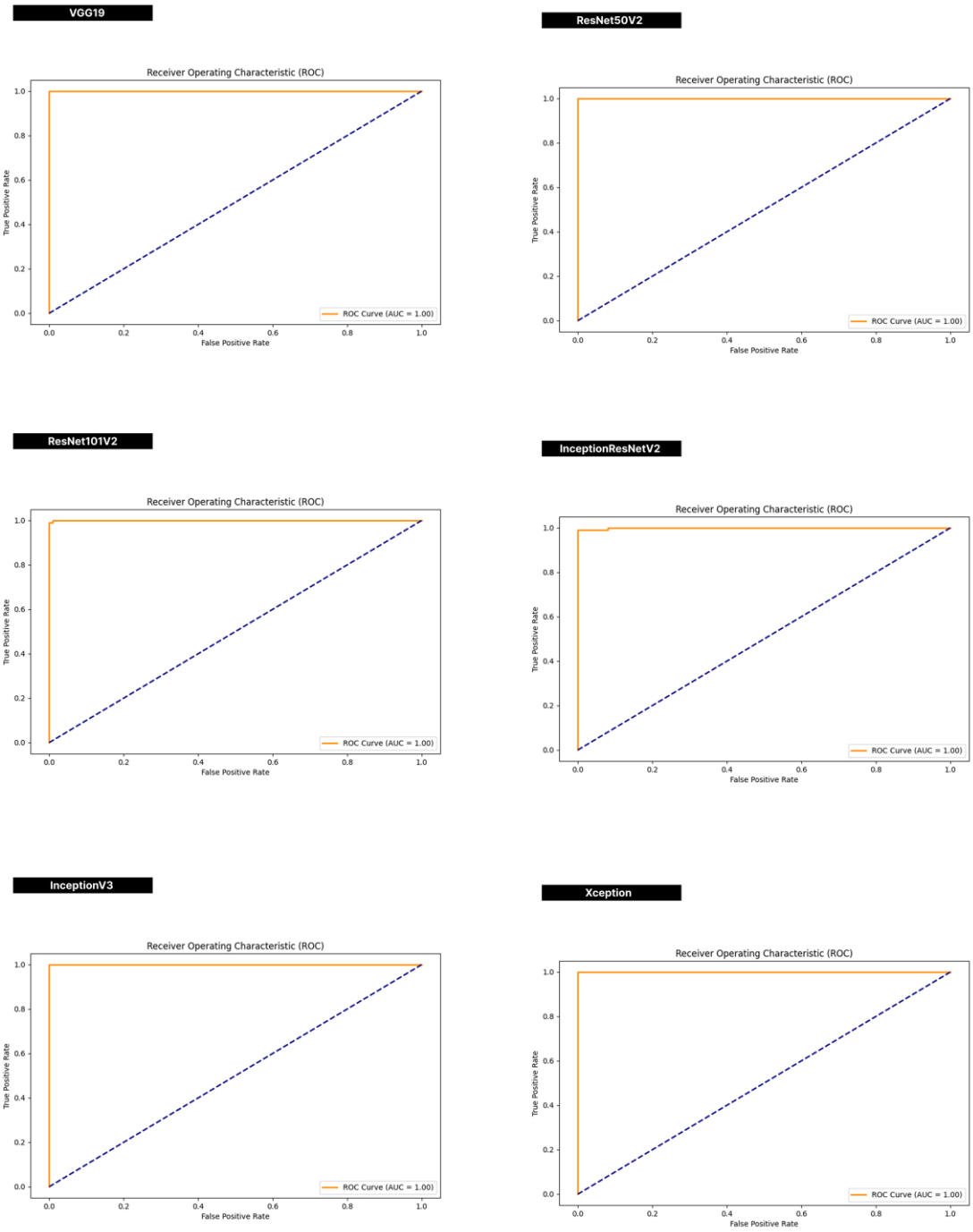


Figure 6: ROC graphs of Models deployed for Binary Classification.

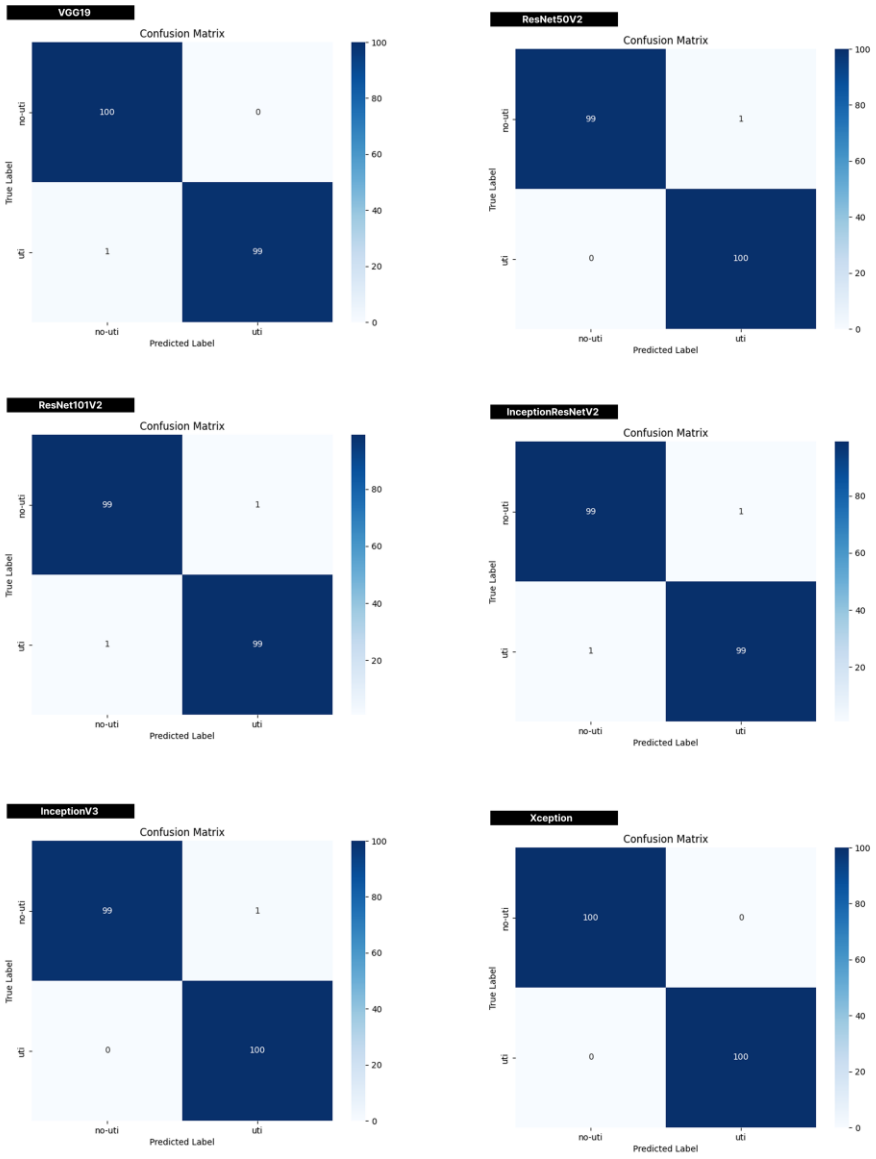


Figure 7: Confusion Matrix of Models deployed for Binary Classification.

4.2 Performance Evaluation of Models Deployed for Multiclass Classification

Multiclass classification or accurate discrimination between different types of urine sediments is crucial for appropriate treatment. Thus, in this study, we deployed several pre-trained CNNs, including InceptionResNetV2, InceptionV3, ResNet50, ResNet101, VGG19 and Xception. The models are evaluated using several metrics, including Precision, Recall, F1-Score, Specificity, MCC, CK, AUC and Test Accuracy.

4.2.1 Performance Evaluation of Models Deployed for Multiclass Classification Using Dataset A

The result of the model performance is presented in Table 6. The accuracy graph of the 6 deployed models is presented in Fig. 8. The ROC graphs are presented in Fig. 9. The results of all deployed models based on 5 rounds are presented in the supplementary file "S2".

Table 6: Performance Evaluation of Models Deployed for Multiclass Classification using Dataset A.

Model Name	Pc	Rc	F1-Score	Sp	AUC	Accuracy	MCC	CK
InceptionResNetV2	90.30	89.37	89.33	94.69	96.90	89.37	84.48	84.06
InceptionV3	86.88	85.28	85.43	92.64	95.00	85.28	78.46	77.91
ResNet50	88.54	86.67	86.86	93.33	94.62	86.67	80.74	79.99
ResNet101	89.32	87.50	87.66	93.75	95.80	87.50	81.99	81.24
VGG19	95.56	94.66	94.72	97.32	99.24	94.66	92.39	91.98
Xception	92.17	91.04	91.08	95.52	97.16	91.04	87.06	86.54

Note: Pc: Precision; Rc: Recall; Sp: Specificity; AUC: Area Under the Curve; Matthews Correlation Coefficient (MCC); Cohen's Kappa (CK).

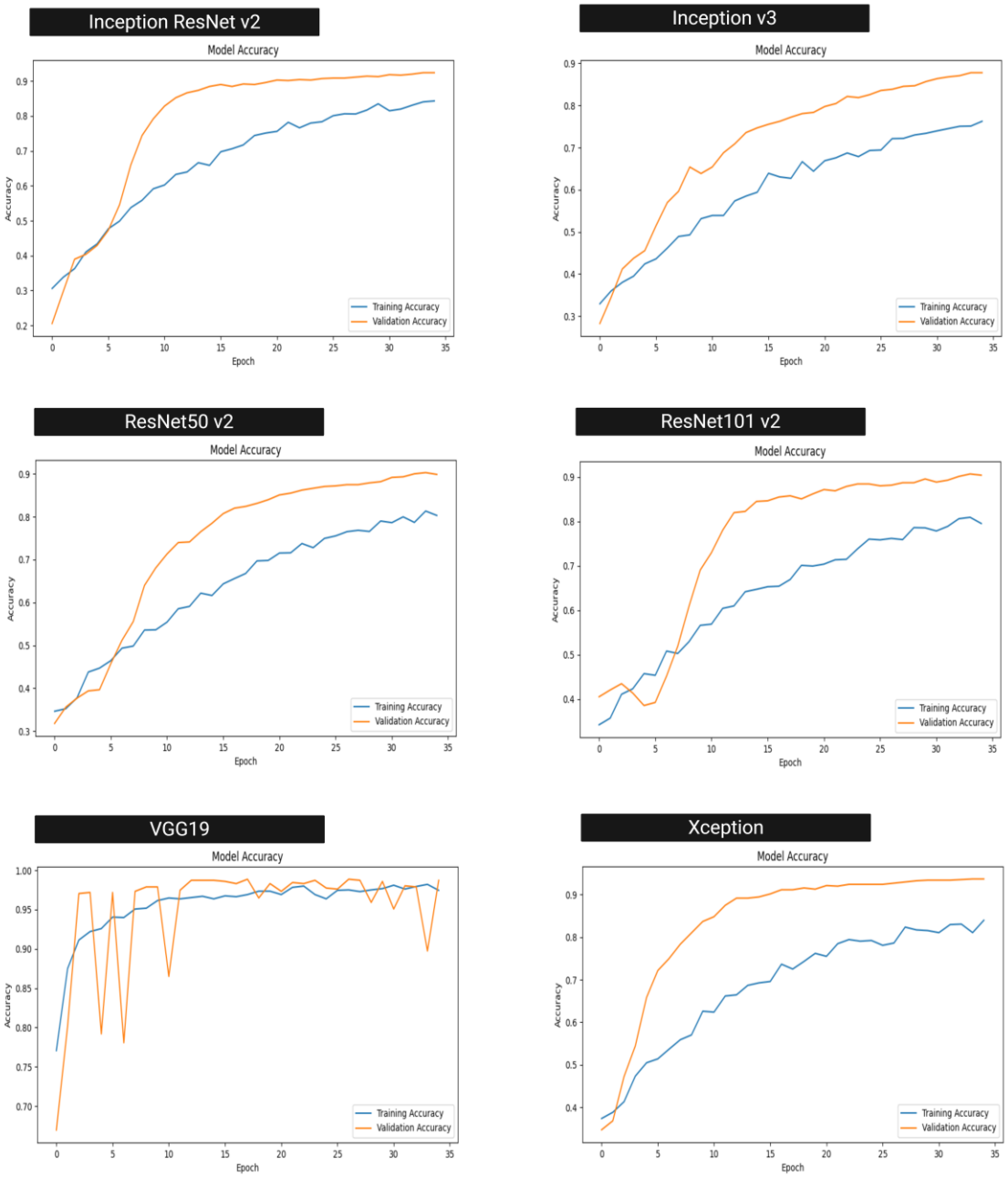


Figure 8: Accuracy graphs of Models deployed for Multiclass Classification Using Dataset A.

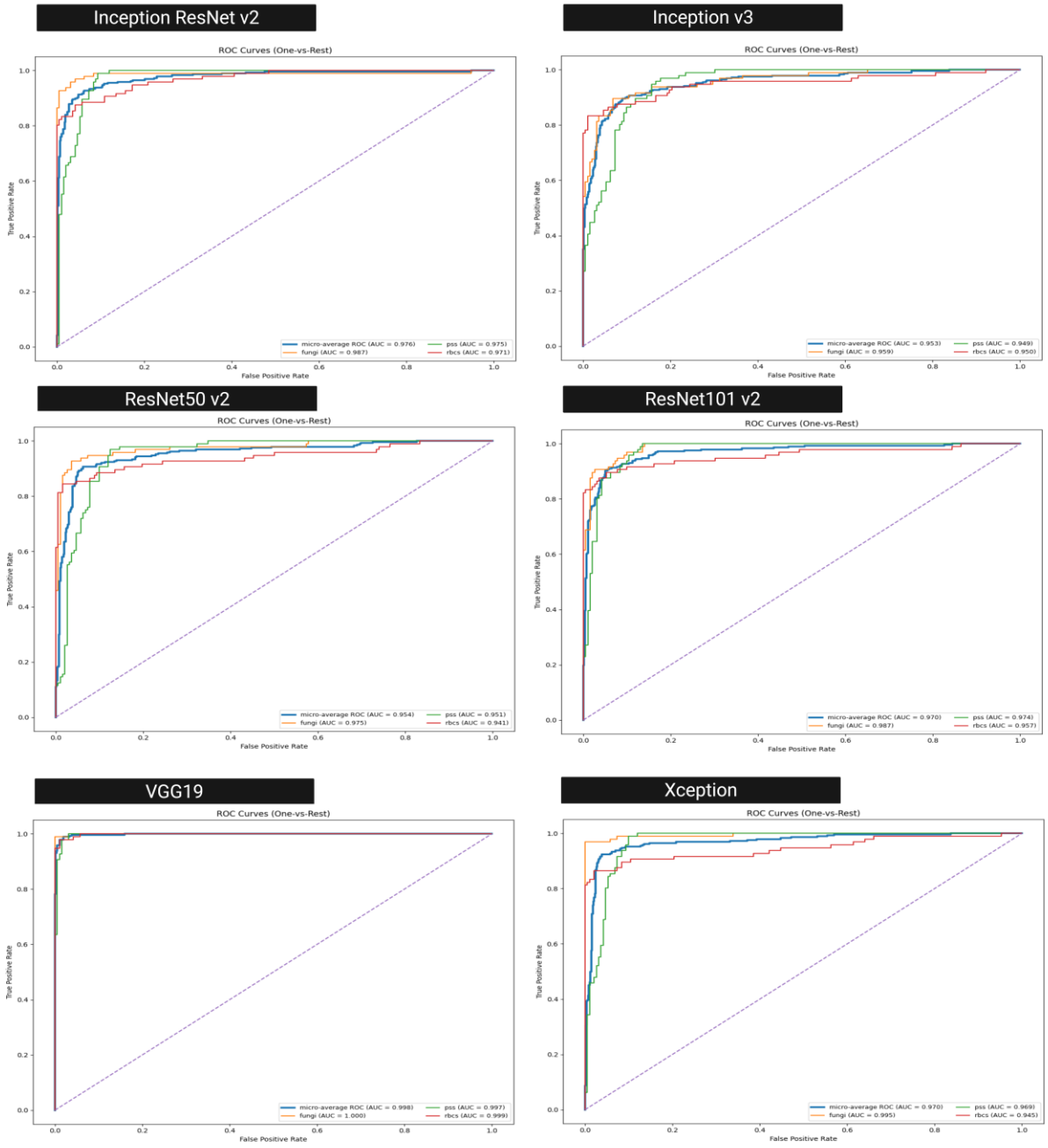


Figure 9: ROC graphs of Models deployed for Multiclass Classification Using Dataset A.

Confusion Matrix of Models Deployed for Multiclass Classification Using Dataset A

In order to test the performance of the deployed models, 288 images (96 fungi, 96 Pus and 96 erythrocytes) are randomly selected. The performance evaluation of the models has shown that InceptionResNetV2 accurately identified 261/288 cases, resulting in 27 miss-classifications. InceptionV3 on the other hand, accurately identified 247/288, resulting in 41 miss-classifications. Consequently, performance evaluation of ResNet50 indicated that the model accurately identified 257/288 cases, resulting in 31 miss-classifications. Subsequently, evaluation of ResNet101 revealed that the model accurately

identified 257/288 cases, resulting in 31 miss-classifications. Likewise, VGG19 accurately identified 282/288 cases, resulting in 6 miss-classifications. Lastly, evaluation of Xception implied that the model accurately identified 266/288 cases, resulting in 22 miss-classifications. The Confusion matrix of models deployed for multiclass classification using dataset A is summarized in Fig. 10.

4.2.2 Performance Evaluation of Models Deployed for Multiclass Classification Using Dataset B

Table 7 presents the performance evaluation of models deployed for Multiclass classification using dataset B. The accuracy graph of the 6 deployed models is presented in Fig. 11. The ROC graphs are presented in Fig. 12. The results of all deployed models based on 5 rounds are presented in the supplementary file “S3”.

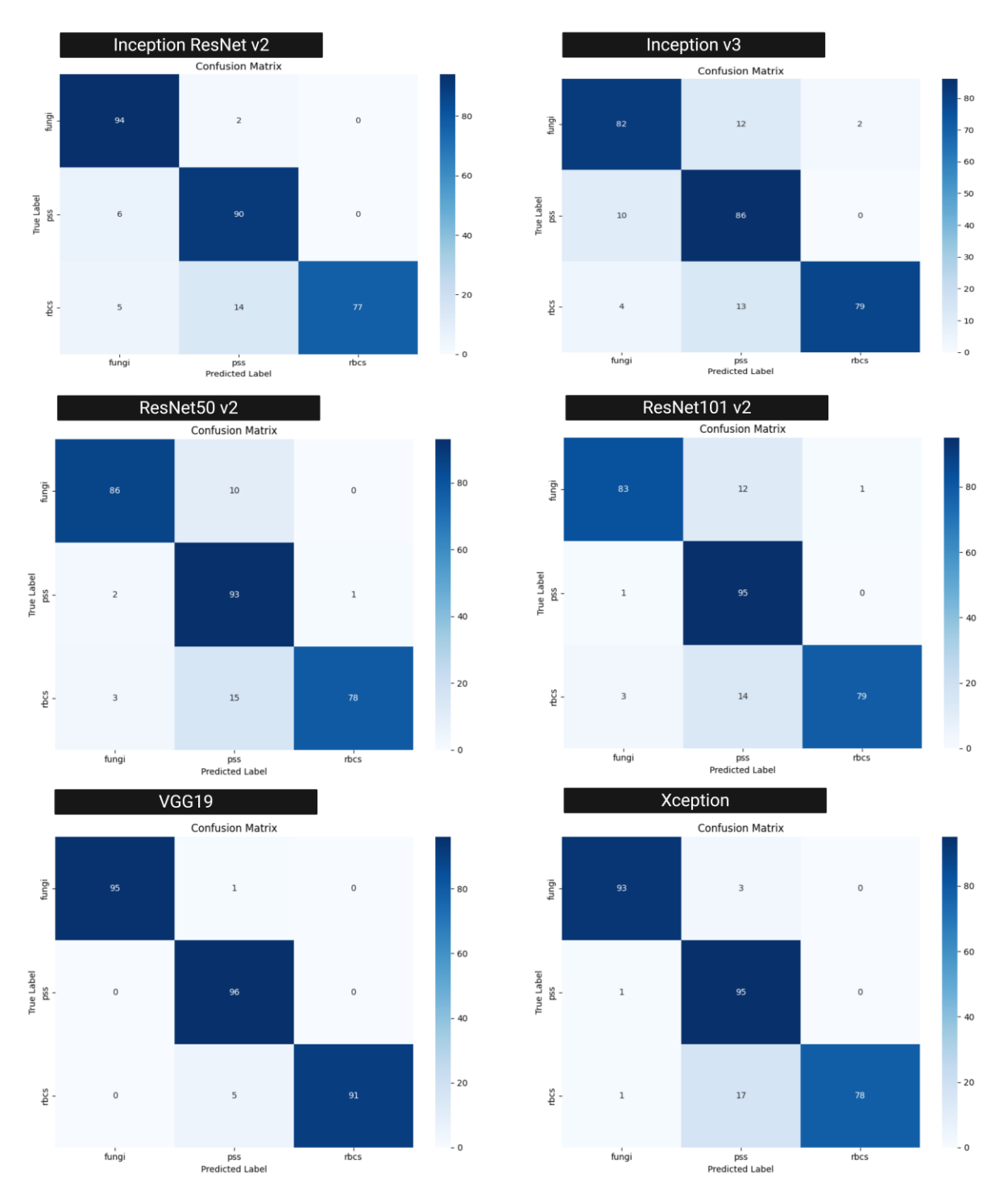


Figure 10: Confusion Matrix of Models deployed for Multiclass Classification Using Dataset A.

Table 7: Performance Evaluation of Models Deployed for Multiclass Classification using Dataset B.

Model Name	Pc	Rc	F1-Score	Sp	AUC	Test Accuracy	MCC	CK
InceptionResNetV2	92.10	85.36	87.00	98.70	99.51	92.72	91.05	90.98

InceptionV3	97.53	97.692	97.60	99.73	99.97	98.46	98.10	98.10
ResNet50	94.16	87.23	88.92	99.09	99.68	94.97	93.80	93.77
ResNet101	97.48	97.15	97.46	99.10	99.95	98.27	97.87	97.87
VGG19	97.41	96.08	96.61	99.64	99.97	97.94	97.47	97.46
Xception	93.88	83.61	84.50	98.81	99.48	93.44	91.92	91.84

Note: Pc: Precision; Rc: Recall; Sp: Specificity; AUC: Area under the Curve; Matthews Correlation Coefficient (MCC); Cohen's Kappa (CK).

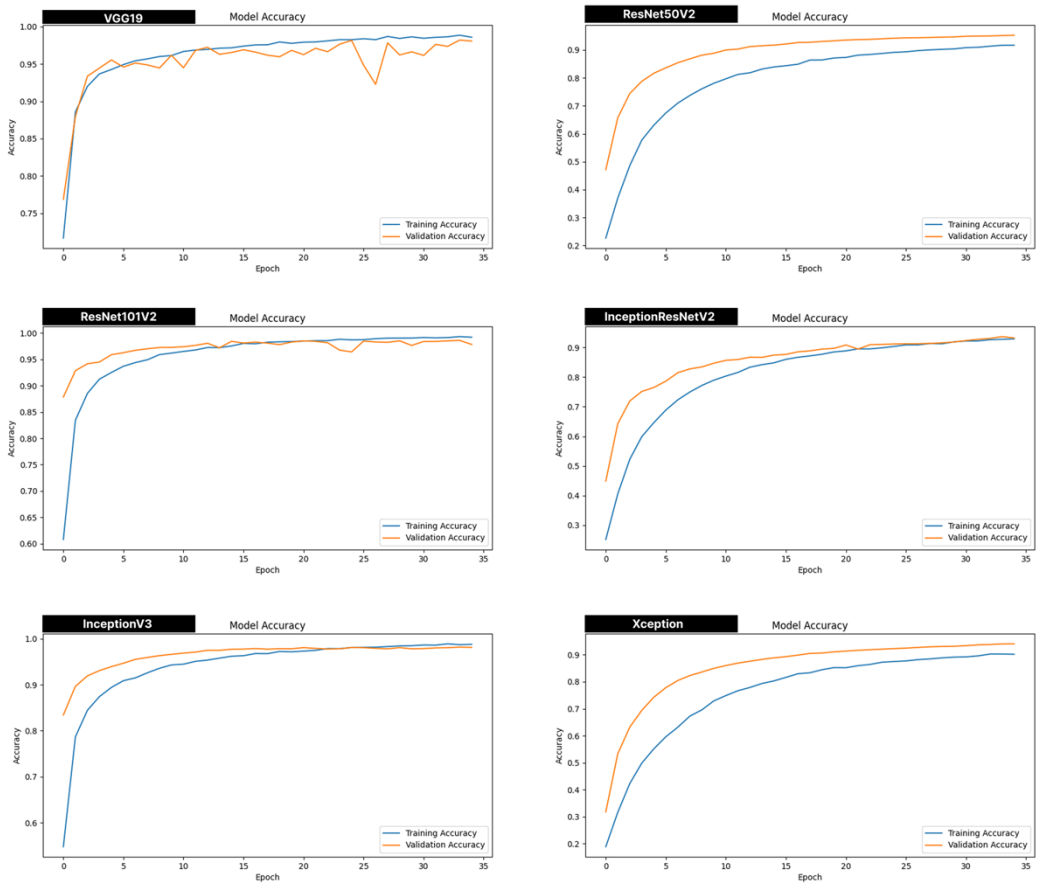


Figure 11: Accuracy graphs of Models deployed for Multiclass Classification Using Dataset B.

Confusion Matrix of Models Deployed for Multiclass Classification Using Dataset B

In order to test the performance of the deployed models, 2649 images (367 casts, 165 crystals, 619 epithelial cells, 70 epithelial nuclei, 600 erythrocytes, 618 leukocytes and 210 mycetes) are randomly selected. The performance evaluation of the models has shown that InceptionResNetV2 accurately identified 359/367 casts, 145/165 crystals, 575/619 epithelial cells, 34/70 epithelial nuclei, 569/600 erythrocytes, 589/618 leukocytes and 201/210 mycetes. InceptionV3, on the other hand, accurately identified 362/367 casts, 162/165 crystals, 610/619 epithelial cells, 64/70 epithelial nuclei, 593/600 erythrocytes, 613/618 leukocytes and 210/210 mycetes. Consequently, performance evaluation of ResNet50 indicated that the model accurately identified 356/367 casts, 154/165 crystals, 597/619 epithelial cells, 31/70 epithelial nuclei, 587/600 erythrocytes, 598/618 leukocytes and 204/210 mycetes.

Subsequently, evaluation of ResNet101 revealed that the model accurately identified 360/367 casts, 164/165 crystals, 606/619 epithelial cells, 68/70 epithelial nuclei, 593/600 erythrocytes, 609/618 leukocytes

and 210/210 mycetes. VGG19 accurately identified 364/367 casts, 164/165 crystals, 615/619 epithelial cells, 58/70 epithelial nuclei, 593/600 erythrocytes, 604/618 leukocytes and all 210 mycetes). Lastly, evaluation of Xception implied that the model accurately identified 352/367 casts, 156/165 crystals, 573/619 epithelial cells, 39/70 epithelial nuclei, 576/600 erythrocytes, 603/618 leukocytes and 205/210 mycetes. The Confusion matrix of models deployed for multiclass classification using dataset B is summarized in Fig. 13.

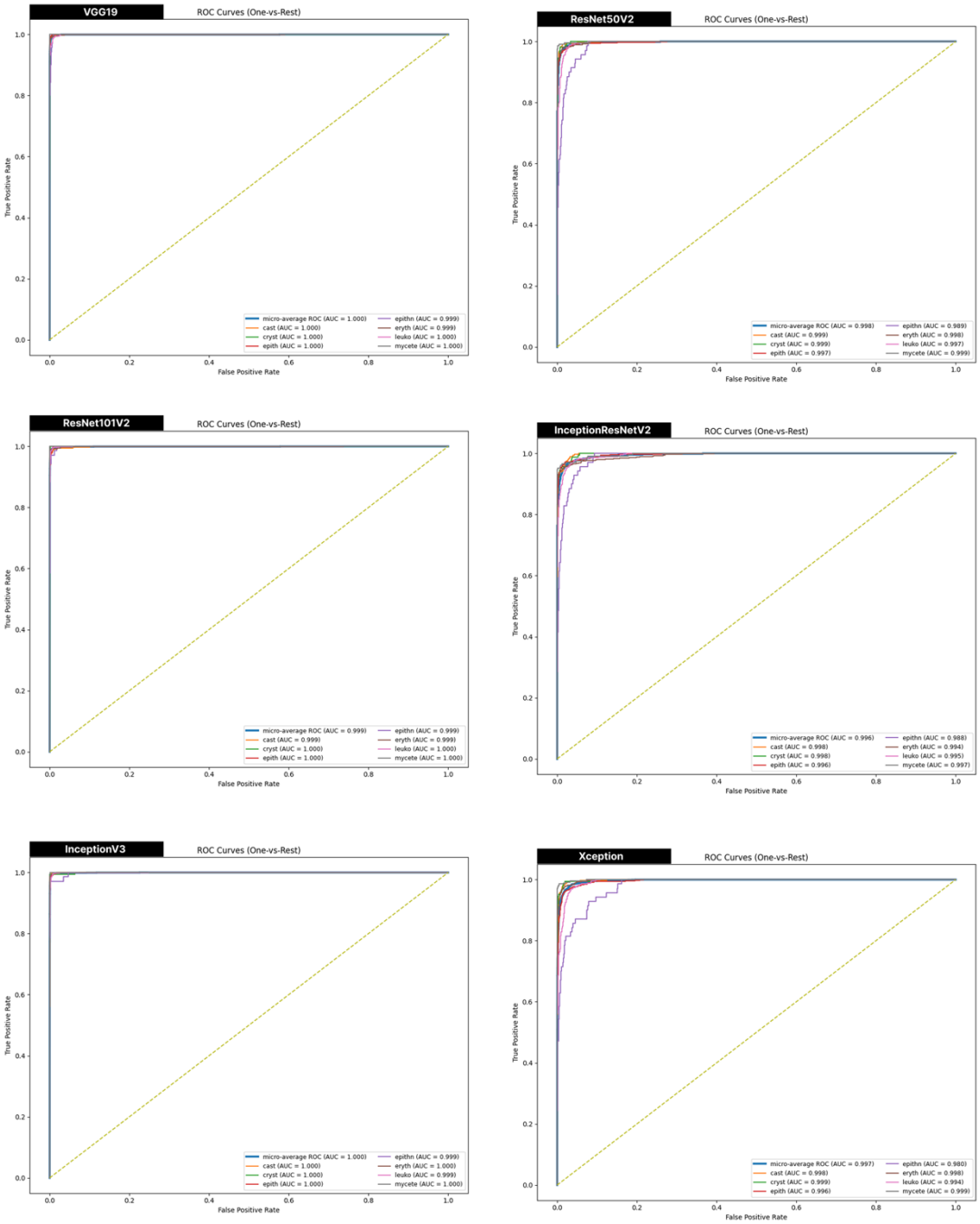


Figure 12: ROC graphs of Models deployed for Multiclass Classification Using Dataset B.

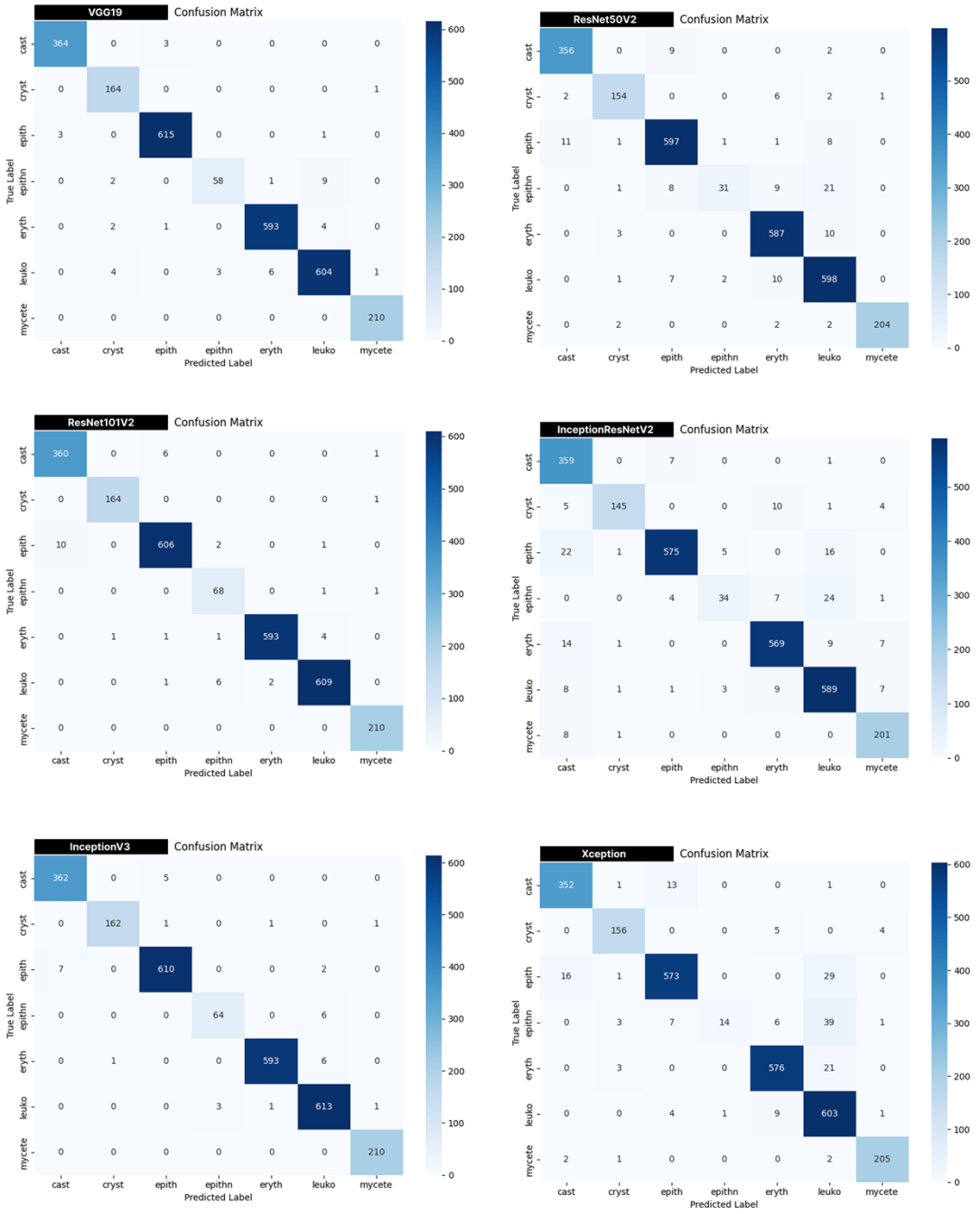


Figure 13: Confusion Matrix of Models deployed for Multiclass Classification Using Dataset B.

4.3 Statistical Analysis

ANOVA is adopted in order to determine if there are any statistically significant differences between the 6 model architectures across 5 training rounds (different random initializations). The results reveal distinct patterns across the three classification tasks.

4.3.1 Binary Classification

For models deployed for binary classification, the results confirm that all the models exhibit very high and similar performance (approximately 98–99% accuracy) and the variability within each model across the five rounds is comparable to the small differences between the models, leading to non-significant ANOVA results (i.e., non-significant differences, $p > 0.05$). This result was expected because binary discrimination between urine sediment and normal cases represents a relatively simple visual task (i.e., the presence versus absence of cellular elements creates distinct feature separability that even shallow and untrained architectures can capture effectively). Thus, Xception's superior results (99.40% accuracy) for binary classification between urine sediments and normal cases stem from its depthwise separable convolutions. By separating spatial and channel-wise convolutions, Xception learns spatial features independently per channel, then combines channels. This decomposition is highly efficient for binary discrimination, where the presence or absence of any cellular material is the primary signal.

4.3.2 Multiclass Dataset A and B

For multiclass using Dataset A and B, the result confirms statistically significant differences ($p < 0.001$) across all performance metrics for the tested models1. Moreover, comparative analysis of deployed models indicates that VGG19 emerged as the best model with a 94.66% mean accuracy using dataset A, and InceptionV3 emerged as the best model with a 98.46% mean accuracy using dataset B. The strongly significant results for both datasets; dataset A ($p = 1.37 \times 10^{-6}$) and dataset B ($p = 1.32 \times 10^{-14}$) confirm an undisputable architectural difference in handling fine-grained sediment classification. The considerably lower p-values for dataset B reflect the increased complexity for discriminating 7 classes, which magnifies architectural strengths and weaknesses. The non-significant "Round" effect ($p = 0.876$) validates experimental consistency, where all the models perform reliably across different initializations, confirming that observed differences stem from architecture rather than stochastic variation.

The exceptional results achieved by VGG19 (94.66% accuracy) on the clinical dataset (i.e., multiclass dataset A) can be attributed to the model's uniform kernel architecture, characterized by the use of 3 × 3 convolutional filters with stride 1 and same padding, which creates a highly regular feature extraction pipeline. Thus, for clinical images with consistent acquisition parameters such as fixed microscope magnification, lighting, and preparation protocol, this regularity enables stable gradient flow and effective learning of subtle morphological distinctions between the 3 classes (fungi, pus cells and erythrocytes).

Subsequently, InceptionV3's top performance (98.46% accuracy) on the diverse public dataset (i.e., multiclass dataset B) can be attributed to the network's multi-scale feature extraction, where the modules apply 1 × 1, 3 × 3, and 5 × 5 convolutions in parallel, thereby capturing features at multiple spatial scales simultaneously. This is critical for dataset B, which contains highly heterogeneous classes (epithelial cells, cast, erythrocytes, crystals, etc.). Secondly, the factorized convolutions of the network factorize larger convolutions into smaller ones (for example, 5 × 5 into two 3 × 3), therefore, reducing parameters while preserving representational capacity, essential for the multiclass problem with limited per-class samples. The result for both binary and multiclass (datasets A and B) is summarized in Table 8.

Table 8: ANOVA Results of Models Deployed for Binary and Multiclass Classification.

Binary			
Metric	**F-Statistic**	***p*-Value**	**Significant**
Test Accuracy	0.9605	0.46	No
AUC	1.2002	0.48	No
Precision	0.8429	0.53	No
Recall	1.1158	0.38	No
F1-Score	0.5827	0.71	No
Specificity	0.8993	0.50	No
MCC	0.9688	0.46	No
Cohen's Kappa	0.9605	0.46	No
Multiclass: Dataset A			
Metric	**F-Statistic**	***p*-Value**	**Significant**
Test Accuracy	14.7889	1.37×10^{-6}	Yes
AUC	38.6039	1.36×10^{-11}	Yes
Precision	21.1488	1.87×10^{-6}	Yes
F1-Score	15.3405	8.82×10^{-7}	Yes
Recall	14.7889	1.37×10^{-6}	Yes
Specificity	14.756	1.30×10^{-6}	Yes
MCC	16.525	2.50×10^{-7}	Yes
Cohen's Kappa	14.795	1.20×10^{-6}	Yes
Multiclass: Dataset B			
Metric	**F-Statistic**	***p*-Value**	**Significant**
Test Accuracy	62.57	1.32×10^{-14}	Yes
AUC	53.22	1.07×10^{-13}	Yes
Precision	3092	8.94×10^{-11}	Yes
F1-Score	44.92	1.68×10^{-12}	Yes
Recall	37.60	1.68×10^{-11}	Yes
Specificity	45.81	1.16×10^{-12}	Yes
MCC	58.42	3.24×10^{-14}	Yes
Cohen's Kappa	58.39	3.28×10^{-14}	Yes

4.4 Real-Time Detection of Urine Sediment: AI/IoT Framework

The real-time urine sediment detection framework operates through three coordinated components, which are the data acquisition layer, the AI inference layer, and the IoT communication layer. Together, they provide instant feedback when predicting urine sediment from microscopic images. At the foundation is a web application that bridges and abstracts the 2 layers. The app functions as a gateway to get input data that can be passed to the AI inference layers using the communication layer as an API.

4.4.1 Technical Implementation: System Architecture

The system is designed using 3 layers. The first layer, known as the data acquisition layer (i.e., client-side), is an interface built with React.js (v18.2.0) and Tailwind CSS. The layer was designed to support the detection of urine sediments using desktop, PC, tablet and mobile devices. The interface enables image upload via drag-and-drop or file selection (format: JPEG/PNG; max size: 10 MB; dimensions: automatically resized to 224 × 224). The second layer, known as IoT Communication (API Gateway), was built with FastAPI (Python 3.9) hosted on AWS EC2 (t3.medium instance). The IoT communication layer enables the system's real-time capabilities by transmitting images and predictions over HTTPS over an API route. While the last layer, known as the AI Inference (Cloud Backend) was designed based on TensorFlow Serving (v2.13.0) containers deployed on AWS ECS, where each model (VGG19 and Inceptionv3) containerized separately for modular maintenance. The response includes predicted class, inference time and model version. The step-by-step procedure for real-time detection is presented in Fig. 14.

4.4.2 Security and Privacy Measures

In order to ensure the confidentiality and integrity of patient data during transmission, all communication between the user-facing web application and the cloud inference server is secured via HTTPS (Hypertext Transfer Protocol Secure). This protocol encrypts the image payload using Transport Layer Security (TLS), thereby preventing unauthorized interception or tampering of the microscopic images as they travel from the point of care to the cloud. Additionally, the API endpoint requires authentication via an API key, which ensures that only authorized instances of the application can submit requests for inference, thus providing a basic layer of access control.

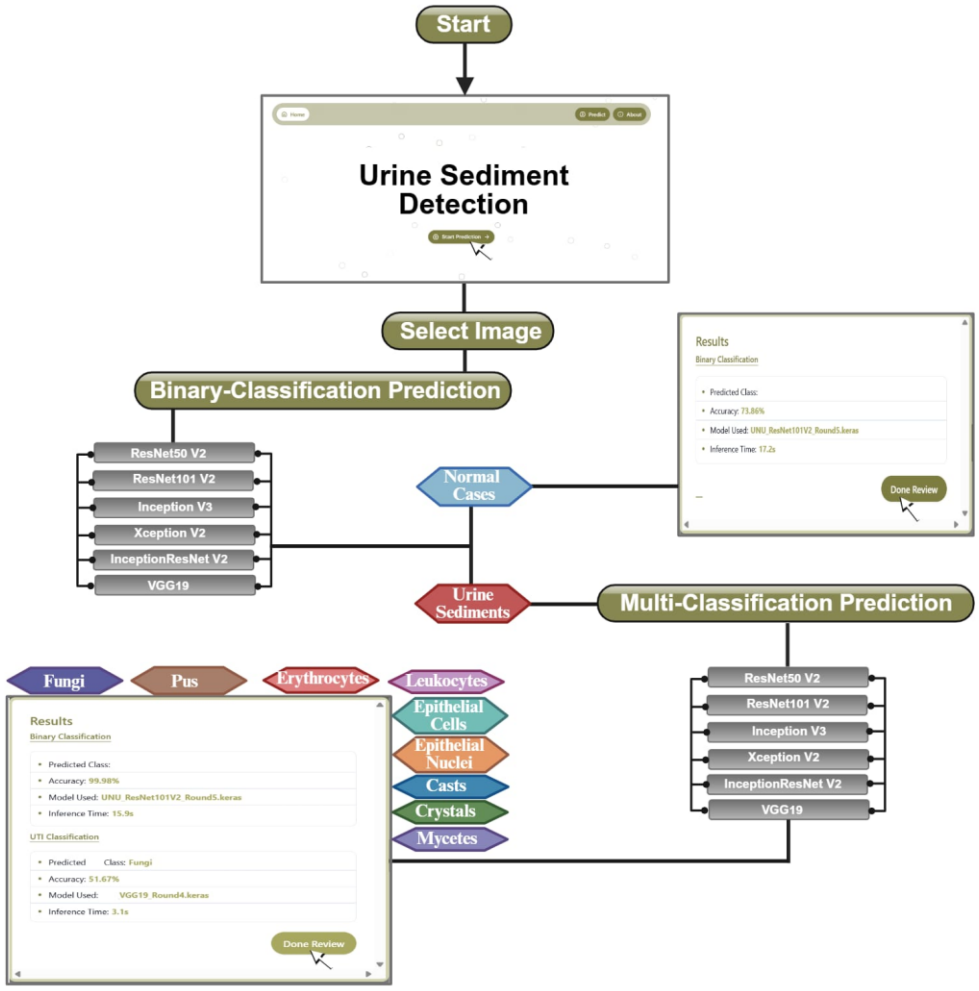


Figure 14: Step-by-step procedure for real-time detection.

4.5 Usability Testing

In order to evaluate the platform's real-world applicability, usability testing was conducted using 15 medical experts (including clinicians and laboratory technologists). The usability testing followed a standardized protocol where each expert was tasked to classify 10 microscopic images of urine sediments (2 from binary classes, 3 from multiclass dataset A and 5 from multiclass dataset B). The tasks performed include uploading microscopic slide images, navigating between binary and multiclass classification modes and comparing the result with ground truths. Performance evaluation of the usability test resulted in 98.7%

(148/150 tasks completed successfully) task success rate and a mean task completion time of 28.4 ± 6.2 s per image.

4.6 Computational Performance Metrics

Computational performance metrics were evaluated for each model by computing average inference time, memory footprint and training convergence as summarized in Table 9. The results indicated that both Xception and InceptionV3 emerged as the most balanced models (i.e., with the lowest memory footprints ≈ 90 MB) and competitive inference times (48–58 ms), therefore, making them highly fitting for real-time IoMT deployment. Despite its high accuracy, VGG19 is the heaviest model (i.e., 550 MB), which may hinder edge deployment. ResNet50, on the other hand, provides the fastest inference time (i.e., 45 ms) and low memory usage, and therefore is ideal for resource-constrained environments. Training convergence is generally faster for binary classification (11–15 epochs) compared to the more complex multiclass tasks (14–22 epochs).

Table 9: Computational Performance Metrics.

Binary			
Models	**Av. Inference Time (ms)**	**Memory Footprint**	**Training Convergence**
VGG19	95	550	14
ResNet50	45	98	12
ResNet101	62	170	13
InceptionV3	48	92	11
Inception ResNetV2	70	210	15
Xception	58	90	12
Multiclass: Dataset A			
Models	**Av. Inference Time (ms)**	**Memory Footprint**	**Training Convergence**
VGG19	95	550	18
ResNet50	45	98	20
ResNet101	62	170	19
InceptionV3	48	92	22
Inception ResNetV2	70	210	21
Xception	58	90	20
Multiclass: Dataset B			
Models	**Av. Inference Time (ms)**	**Memory Footprint**	**Training Convergence**
VGG19	95	550	16
ResNet50	45	98	20
ResNet101	62	170	15
InceptionV3	48	92	14
Inception ResNetV2	70	210	18
Xception	58	90	22

4.7 Real-Time Inference Latency

Considering the fact that in real-world IoMT scenarios, input quality may vary, there is a need to assess model robustness and reliability. Thus, a numerical stability test was conducted. This test is critical for IoMT deployment, as it highlights which models are more likely to maintain performance under suboptimal imaging conditions such as sensor noise and compression artifacts. The test is conducted by systematically adding Gaussian noise ($\sigma = 0.05$) to the test images, and the degradation in classification accuracy was subsequently measured.

The result of the analysis indicated that ResNet50 demonstrated the best overall stability for binary classification, resulting in the smallest accuracy drop of 2.18%. For multiclass classification using dataset A, the result indicated that VGG19 demonstrated the greatest robustness with only a 3.86% drop. While for multiclass classification using dataset B, the result indicated that InceptionV3 demonstrated the greatest

robustness with only a 2.36% drop. The summary of the numerical stability analysis for binary and multiclass classifications (A and B) is presented in Table 10.

Table 10: Real-Time Inference Latency.

Binary			
Models	**Original Test Accuracy (%)**	**Accuracy with Noise (%)**	**Accuracy Drop (%)**
VGG19	98.60	96.40	2.20
ResNet50	98.88	96.70	2.18
ResNet101	98.40	96.10	2.30
InceptionV3	99.00	96.80	2.20
Inception ResNetV2	98.70	96.30	2.40
Xception	99.40	97.20	2.20
Multiclass: Dataset A			
Models	**Original Test Accuracy (%)**	**Accuracy with Noise (%)**	**Accuracy Drop (%)**
VGG19	94.66	90.80	3.86
ResNet50	88.67	82.40	4.27
ResNet101	87.50	83.10	4.40
InceptionV3	85.28	80.90	4.38
Inception ResNetV2	89.37	85.20	4.17
Xception	91.04	87.10	3.94
Multiclass: Dataset B			
Models	**Original Test Accuracy (%)**	**Accuracy with Noise (%)**	**Accuracy Drop (%)**
VGG19	97.94	95.20	2.74
ResNet50	94.97	92.10	2.87
ResNet101	98.27	95.80	2.47
InceptionV3	98.46	96.10	2.36
Inception ResNetV2	92.72	89.50	3.22
Xception	93.44	90.20	3.24

4.8 Discussion

A UTI is characterized as an infection that affects the urinary system and associated organs such as the Kidneys, bladder, urethra and ureters [24]. UTIs include Pyelonephritis, Cystitis, Urethritis, etc. In order to diagnose patients suspected of UTIs, medical experts rely on several techniques including, urinalysis, Urine Culture and Sensitivity (C&S) [25]. Urinalysis is one of the common conventional techniques used by medical experts for the detection of UTIs. Despite its reliance, the technique is hampered by several limitations, which include false positive and negative results, time-consuming, technician-dependent and the probability of contamination. Moreover, manual microscopic examination is variable and requires a skilled technician [8,25].

In order to address these challenges, scientists integrate AI-based techniques. AI-based techniques, particularly ML and DL, offer powerful tools to overcome the bottlenecks and inaccuracies in traditional UTI diagnosis and prediction [20,23,26]. Several ML and DL models have been trained on vast datasets containing urinalysis results (such as dipstick readings, WBC counts, etc.) and their corresponding culture results. These models learn complex, non-linear patterns that humans might miss, resulting in high accuracy and sensitivity. Moreover, DL models, CNNs, have been deployed to analyze digital images of urine sediment [27,28]. These models have been shown to automatically identify, count, and classify RBCs, WBCs, epithelial cells, casts, and, crucially, bacteria [22, 27, 28].

Despite the numerous studies on the topic, several existing studies rely on data curated from public repositories. Moreover, the majority of existing studies either deployed DL-based models for multiclass classification. However, this study deployed several state-of-the-art DL pre-trained CNN models, including InceptionResNetV2, InceptionV3, ResNet50, ResNet101, VGG19, and Xception for the identification of urine sediments, using a dataset acquired from a hospital and a public repository. The pretrained models deployed were assessed on two critical tasks: binary classification (discriminating urine sediments from normal cases) and multiclass classification (identifying specific types of urine sediments, such as fungi, pus, and erythrocytes, cast, mycetes, etc.).

The binary classification task is necessary for initial screening, aiming to accurately distinguish between normal and urine sediment cases. Evaluation of the performance of all six deployed pre-trained CNN models presents an exceptional result, in which all the models achieved an average test accuracy above 97%. This indicates a high level of proficiency across the board for this task. Comparative analysis of the deployed models for binary classification, as summarized in Table 4, showcases Xception as the top-performing model, achieving the highest scores across 5/6 metrics, including precision (99.21%) and test accuracy (99.40%). Subsequently, InceptionV3 ranked second, achieving a perfect Recall of 100% (i.e., correctly identified all the urine sediment cases), an essential characteristic for a medical screening tool to avoid missing true cases (false negatives). Moreover, the model achieved a significant accuracy of 99.00%.

The subsequent comparative analysis of the performance of the models based on confusion matrices also reflects the result achieved on the test set, as summarized in Fig. 7. Complimenting the result achieved on the test set, Xception identified all the 200 images (100 urine sediments and normal cases), directly correlating with their high accuracy and balanced precision-recall scores. While both InceptionV3, ResNet50 and VGG19 identified 199/200 images.

Compared with binary classification, multiclass classification of urine sediments is more complex as it requires AI-based models to not only detect features but also to categorize their type, which is crucial for determining the appropriate type of UTI. As expected, the performance of all deployed models was lower than that of binary classification, yet several models still produced outstanding results. Evaluation of the performance of all six deployed pre-trained CNN models trained using dataset A (as shown in Table 6) indicated that VGG19 emerged as the superior model, achieving the highest scores across all six metrics, with an accuracy of 94.66% and an exceptional AUC of 99.24%. This result highlights its capability to handle fine-grained distinctions between multiple urine sediments. Xception held a clear second place, being the only other model to exceed 91% across 6 metrics (accuracy, recall, precision, F1-score, specificity,

and AUC). This demonstrates its strong generalizability for complex classification problems. Consequently, InceptionV3 remained the least performant model with an accuracy of 85.28%.

The following comparative analysis of the performance of the models based on confusion matrices also reflects the result achieved on the test set, as summarized in Table 6. As shown in Fig. 10, VGG19 demonstrates its capability in discriminating between 3 subtypes of urine sediments, with a remarkable 282 correct identifications out of 288 total cases, resulting in only 6 errors. This directly translates to its top-tier metrics. Xception retain second-place ranking is supported by its 266 correct identifications. The higher miss-classifications of InceptionV3 (i.e., 41) align with its lower scores across performance metrics.

Consequently, the performance evaluation of the 6 deployed models for the discrimination between 7 subtypes of urine-sediments (i.e., dataset B), as shown in Table 7, indicated that InceptionV3 emerged as the superior model across all the 8 metrics, achieving an accuracy of 98.46%. ResNet50 ranked second with 98.27% accuracy, while VGG19 ranked third with 97.94%.

The following comparative analysis of the performance of the models based on confusion matrices also reflects the result achieved on the test set, as summarized in Table 7. As shown in Fig. 13, both InceptionV3 and VGG19 demonstrated their capability in discriminating between subtypes of urine sediments, resulting in fewer miss-classifications. While InceptionResNetV2 and Xception performed poorly, this reflects their performance on the test sets.

4.9 Comparison with Related Work

Considering the fact that existing studies focus on multiclass classifications of urine sediment into 7 or more categories, to ensure realistic comparative analysis, only the result based on multiclass classification using dataset B is compared with existing studies. Among several models trained using the multiclass dataset B, InceptionV3 emerged as the superior model across all the metrics (i.e., 97.94% accuracy, 97.53% mAP and 99.97% AUC). Based on accuracy, InceptionV3outperforms the study reported by [22] (AlexNet-based Faster R-CNN: 98.6% accuracy), [23] (RetinaNet: 88.65% accuracy) and [17] (YUS-Net: 96.07% accuracy). Consequently, InceptionV3 also emerged as the best model based on mAP (97.53%), outperforming the models deployed by [17] (YUS-Net: 96.07% mAP), [7] (Ensemble YOLOv9e + KD-YOLOX-ViT: 94.18% mAP), [20] (YOLOv5: 85.8% mAP) and [21] (DFPN: 86.9% mAP). Comparison with related work is presented in Table 11.

Table 11: Comparison with related work.

Ref.	Dataset + Size	# Classes	Models	Result
[7]	USE (5376 images)	7	Ensemble (YOLOv9e and KD-YOLOX-ViT)	94.18% mAP
[17]	USE (5376 images)	7	YUS-Net	96.07% accuracy and 96.07 mAP
[18]	Urised11 (7364 images)	11	DP-YOLO	41.2% AP on COCO2017 and 49.2% AP on Urised11
[19]	300 images (3562 annotated cells)	7	Patch U-Net	Precision of 0.890, Recall of 0.920, and an AUC of 0.989.
[20]	5376 images	6	YOLOv5	85.8% mAP @ IoU 0.5
[21]	USE dataset (5377 images)	7	DFPN	mAP of 86.9%.
[22]	500 images (51,077 annotated patches)	10	AlexNet-based Faster R-CNN	98.6% accuracy
[23]	15,360 images	7	RetinaNet	88.65% accuracy
This study	USE (5376 images)	7	InceptionV3	97.94% accuracy, 97.53% mAP and 99.97% AUC

4.10 Limitations and Future Work

Even though this study achieved significant results in terms of discrimination between urine sediment and normal cases and different subtypes of urine sediments, the work is not without limitations. One of

these limitations is the deployment of a few models such as VGG19, Inceptionv3, ResNet101 and ResNet50. Thus, future studies will explore other state-of-the-art models such as DenseNet, EfficientNet, Transformer-based models such as DTSNet [29], ensemble models, hybrid models, etc. Consequently, this study only acquired two datasets (clinical and online repository), limited to 7 classes. Thus, future study will explore other clinical and publicly accessible datasets. Another limitation of this study is the trade-offs associated with the class balancing strategy employed for the multiclass dataset A, where the erythrocytes and pus classes were downsampled to match the size of fungi (the smallest class). Even though this strategy effectively eliminates class bias during training and simplifies model optimization, it intrinsically results in the omission of a substantial number of potentially informative samples from the original dataset. Therefore, in order to mitigate this loss, future studies will explore alternative strategies, such as employing more sophisticated augmentation techniques for the minority class.

5 Conclusion

The precise detection of urinary particles is crucial for diagnosing UTIs and related conditions, such as kidney disease (nephropathy), in clinical urinalysis. The current standard practice of microscopy, which the healthcare sector relies on, is prone to human error, subjective, labor-intensive and time-consuming. Existing studies have attempted to bridge this gap using AI-based techniques. However, many existing particle detection algorithms lack adequate accuracy and are restricted to offline analysis. This study proposes an AI/IoT method to overcome these limitations. The overall methodology revolves around a dual-task framework that performs binary classification (urine sediment vs. normal cases) and 2 multiclass classifications using dataset A (fungi, pus, erythrocytes) and dataset B (casts, crystals, erythrocytes leukocytes, epithelial cells, epithelial nuclei, mycetes). Several pre-trained CNNs are deployed, including VGG19, ResNet50, ResNet101, Inception, InceptionResNet and Xception, trained and tested using both the hospital-acquired dataset (3372 images) and USE (5377 images and 26,419 cropped microscopic images). Evaluation of the performance and comparative assessment of all the 6 deployed pre-trained CNN models showcase Xception as the top-performing model for binary classification, VGG19 for multiclass classification using dataset A and InceptionV3 for multiclass classification using dataset B. Despite achieving satisfactory results, the overall framework can be improved by acquiring more clinical datasets, validation using publicly accessible domains, extensive and rigorous data augmentation, model optimization, and ensemble learning.

Acknowledgement: We will like to acknowledge BEN RUCHD Laboratory, Tobruk Libya for sharing the clinical data used in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Abdullahi Umar Ibrahim, Suleiman Asir and Fadi Al-Turjman; methodology, Abdullahi Umar Ibrahim, Mohamed Ahmed Mohamed Ahmed, Ibrahim Ahmed Ame and Chidi Wilson Nwekwo; formal analysis, Fathy A A. Hassan, Samuel Nii Tackie and John Bush Idoko; data curation, Mohamed Ahmed Mohamed Ahmed and Fathy A. A. Hassan; writing—original draft preparation, Abdullahi Umar Ibrahim and Ibrahim Ahmed Ame; writing—review and editing, John Bush Idoko, Suleiman Asir and Fadi Al-Turjman; supervision, Suleiman Asir and Fadi Al-Turjman. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, [AUI], upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Supplementary Materials: The supplementary material is available online at https://doi.org/10.23967/j.rimni.2026.10.78308.

References

1. Center for Disease Control and Prevention (CDC). Urinary tract infection basics. [cited 2025 Aug 28]. Available from: https://www.cdc.gov/uti/about/index.html.
2. Medina M, Castillo-Pino E. An introduction to the epidemiology and burden of urinary tract infections. Ther Adv Urol. 2019;11:1756287219832172. doi:10.1177/1756287219832172.
3. Zeng Z, Zhan J, Zhang K, Chen H, Cheng S. Global, regional, and national burden of urinary tract infections from 1990 to 2019: an analysis of the global burden of disease study 2019. World J Urol. 2022;40(3):755–63. doi:10.1007/s00345-021-03913-0.
4. Yang X, Chen H, Zheng Y, Qu S, Wang H, Yi F. Disease burden and long-term trends of urinary tract infections: a worldwide report. Front Public Health. 2022;10:888205. doi:10.3389/fpubh.2022.888205.
5. He Y, Zhao J, Wang L, Han C, Yan R, Zhu P, et al. Epidemiological trends and predictions of urinary tract infections in the global burden of disease study 2021. Sci Rep. 2025;15(1):4702. doi:10.1038/s41598-025-89240-5.
6. Li L, Li Y, Chen Y, Hou H, Wang J, Liu M, et al. Global, regional, and national lifetime probabilities of urinary tract infections and interstitial nephritis from 1990 to 2021. J Health Popul Nutr. 2025;44(1):231. doi:10.1186/s41043-025-00950-y.
7. Naznine M, Salam A, Khan MM, Nobi SF, Chowdhury MEH. An ensemble deep learning approach for accurate urinary sediment detection using YOLOv9e and KD-YOLOX-ViT. IEEE Access. 2025;13:101934–54. doi:10.1109/ACCESS.2025.3577959.
8. Schmiemann G, Kniehl E, Gebhardt K, Matejczyk MM, Hummers-Pradier E. The diagnosis of urinary tract infection: a systematic review. Dtsch Arztebl Int. 2010;107(21):361–7. doi:10.3238/arztebl.2010.0361.
9. Schulz L, Hoffman RJ, Pothof J, Fox B. Top ten myths regarding the diagnosis and treatment of urinary tract infections. J Emerg Med. 2016;51(1):25–30. doi:10.1016/j.jemermed.2016.02.009.
10. Xie R, Li X, Li G, Fu R. Diagnostic value of different urine tests for urinary tract infection: a systematic review and meta-analysis. Transl Androl Urol. 2022;11(3):325–35. doi:10.21037/tau-22-65.
11. Mouraviev V, McDonald M. An implementation of next generation sequencing for prevention and diagnosis of urinary tract infection in urology. Can J Urol. 2018;25(3):9349–56.
12. Zhao M, Qi S, Sun Y, Zheng X. Comparison of polymerase chain reaction and next-generation sequencing with conventional urine culture for the diagnosis of urinary tract infections: a meta-analysis. Open Med. 2024;19(1):20240921. doi:10.1515/med-2024-0921.
13. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med. 2020;7:27. doi:10.3389/fmed.2020.00027.
14. Irkham I, Ibrahim AU, Nwekwo CW, Al-Turjman F, Hartati YW. Current technologies for detection of COVID-19: biosensors, artificial intelligence and internet of Medical Things (IoMT): review. Sensors. 2022;23(1):426. doi:10.3390/s23010426.
15. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. BMC Med Inform Decis Mak. 2019;19(1):171. doi:10.1186/s12911-019-0878-9.
16. De Bruyne S, De Kesel P, Oyaert M. Applications of artificial intelligence in urinalysis: is the future already here? Clin Chem. 2023;69(12):1348–60. doi:10.1093/clinchem/hvad136.
17. Lyu H, Xu F, Jin T, Zheng S, Zhou C, Cao Y, et al. Automated detection of multi-class urinary sediment particles: an accurate deep learning approach. Biocybern Biomed Eng. 2023;43(4):672–83. doi:10.1016/j.bbe.2023.09.003.
18. Wang C, Wang Q, Qian Y, Hu Y, Xue Y, Wang H. DP-YOLO: effective improvement based on YOLO detector. Appl Sci. 2023;13(21):11676. doi:10.3390/app132111676.
19. Liou N, De T, Urbanski A, Chieng C, Kong Q, David AL, et al. A clinical microscopy dataset to develop a deep learning diagnostic test for urinary tract infection. Sci Data. 2024;11(1):155. doi:10.1038/s41597-024-02975-0.
20. Suhail K, Brindha D. Microscopic urinary particle detection by different YOLOv5 models with evolutionary genetic algorithm based hyperparameter optimization. Comput Biol Med. 2024;169:107895. doi:10.1016/j.compbiomed.2023.107895.
21. Liang Y, Tang Z, Yan M, Liu J. Object detection based on deep learning for urine sediment examination. Biocybern Biomed Eng. 2018;38(3):661–70. doi:10.1016/j.bbe.2018.05.004.
22. Avci D, Sert E, Dogantekin E, Yildirim O, Tadeusiewicz R, Plawiak P. A new super resolution Faster R-CNN model based detection and classification of urine sediments. Biocybern Biomed Eng. 2023;43(1):58–68. doi:10.1016/j.bbe.2022.12.001.

23. Li Q, Yu Z, Qi T, Zheng L, Qi S, He Z, et al. Inspection of visible components in urine based on deep learning. Med Phys. 2020;47(7):2937–49. doi:10.1002/mp.14118.
24. Mancuso G, Midiri A, Gerace E, Marra M, Zummo S, Biondo C. Urinary tract infections: the current scenario and future prospects. Pathogens. 2023;12(4):623. doi:10.3390/pathogens12040623.
25. Chu CM, Lowder JL. Diagnosis and treatment of urinary tract infections across age groups. Am J Obstet Gynecol. 2018;219(1):40–51. doi:10.1016/j.ajog.2017.12.231.
26. Capstick A, Palermo F, Zakka K, Fletcher-Lloyd N, Walsh C, Cui T, et al. Digital remote monitoring for screening and early detection of urinary tract infections. npj Digit Med. 2024;7(1):11. doi:10.1038/s41746-023-00995-5.
27. Suhail K, Brindha D. A review on various methods for recognition of urine particles using digital microscopic images of urine sediments. Biomed Signal Process Control. 2021;68:102806. doi:10.1016/j.bspc.2021.102806.
28. 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. Comput Meth Biomech Biomed Eng Imag Vis. 2020;8(1):109–14. doi:10.1080/21681163.2019.1608307.
29. Xiao W, Li X, Hu L, Hao Y, Chen M. DTSNet: dynamic transformer slimming for efficient vision recognition. IEEE Trans Multimed. 2026;28:1589–600. doi:10.1109/TMM.2025.3607796.