Afro-Asian Journal of Scientific Research (AAJSR)
المجلة الأفروآسيوية للبحث العلمي
E-ISSN: 2959-6505
Volume 4, Issue 2, 2026
Page No: 12-19
Website: https://aajsr.com/index.php/aajsr/index
SJIFactor 2024: 5.028 ISI 2025: 0.915 معامل التأثير العربي (AIF) 2025: 0.76

Deep Learning Architectures Comparison for Kidney Stone Classification in CT Images

Abdulhalim Alahrash[1*], Aeyman Hassan[2]
[1,2]Department of Computer Engineering, University of Zawia, Zawia, Libya

مقارنة بين هياكل التعلم العميق لتصنيف حصى الكلى في صور الأشعة المقطعية

عبدالحليم الاحرش[1*]، أيمن حسين[2]
[1,2]قسم هندسة الحاسوب، كلية الهندسة، جامعة الزاوية، الزاوية، ليبيا

*Corresponding author: a.hassan@zu.edu.ly

Received: January 12, 2026 | Accepted: February 16, 2026 | Published: February 26, 2026

Abstract:
Kidney stones are a common condition that may cause severe pain and complications if not diagnosed early. Computed tomography (CT) imaging is widely used for identifying renal stones due to its high sensitivity. This study focuses on binary image-level classification (stone vs. normal) in CT images using deep learning architectures. Three convolutional neural networks—ResNet50, XResNet50, and DenseNet201—were evaluated under a unified preprocessing and training pipeline. Two publicly available CT datasets were used, each divided into 70% training and 30% testing sets. Models were trained using Google Colab with data augmentation to reduce overfitting. Performance was assessed using accuracy, precision, recall, and F1-score. Across the two datasets, XResNet50 achieved the highest accuracy (97% and 92%, respectively). While the reported results indicate strong performance under the defined experimental setup, further validation using patient-wise splitting and additional datasets is recommended to confirm generalisation. These findings provide a comparative reference for selecting suitable architectures for CT-based kidney stone classification.

Keywords: CT Images, Deep Learning, Kidney Stones.

الملخص
حصى الكلى حالة شائعة قد تسبب ألمًا شديدًا ومضاعفات إذا لم يتم تشخيصها مبكرًا. يُستخدم التصوير المقطعي المحوسب (CT)على نطاق واسع لاكتشاف حصي الكلي نظرًا لحساسيته العالية. تركز هذه الدراسة على التصنيف الثنائي على مستوى الصورة (حصوة مقابل طبيعي) في صور الأشعة المقطعية باستخدام معماريات التعلم العميق. تم تقييم ثلاث شبكات عصبية التفافية وهي ResNet50 و XResNet50 و DenseNet201، ضمن إطار موحد لعمليات المعالجة المسبقة والتدريب. استُخدمت مجموعتا بيانات عامتان لصور الأشعة المقطعية، حيث قُسمت كل مجموعة إلى 70% للتدريب و30% للاختبار. تم تدريب النماذج باستخدام Google Colab مع تطبيق تقنيات زيادة البيانات بأستخدام مايعرف ب (Data Augmentation)لتقليل فرط التكيّف (Overfitting).تم تقييم الأداء باستخدام مقاييس الدقة (Accuracy) ، والدقة الإيجابية (Precision) ، والاسترجاع(Recall) ، ودرجة F1 عبر مجموعتي البيانات، حقق نموذج XResNet50أعلى دقة بلغت 97% و92% على التوالي.ورغم أن النتائج تشير إلى أداء قوي ضمن الإعداد التجريبي المحدد، يُوصى بإجراء تحقق إضافي باستخدام تقسيم البيانات على مستوى المرضى (Patient-wise Splitting) ومجموعات بيانات إضافية لتأكيد قابلية التعميم.تقدم هذه النتائج مرجعًا مقارنًا لاختيار المعماريات المناسبة لتصنيف حصى الكلى اعتمادًا على صور الأشعة المقطعية.

الكلمات المفتاحية: صور الأشعة المقطعية(CT) ، التعلم العميق، حصى الكلى.

Introduction

DNNs have shown significant growth due to their capabilities in various fields. They have played a critical role in several medical applications. Convolutional Neural Networks (CNNs) are an example of DNNs that have enhanced the detection accuracy of several diseases through CT medical images. ResNet, DenseNet, and EfficientNet are deep learning models with distinct architectures evaluated recently for kidney stone detection [1]. In this study, the DenseNet model achieved an accuracy of 86%, which outperformed the other deep learning models. Similarly, lightweight architectures such as MobileNet improved performance while reducing computational cost. In 2023, MobileNet and depthwise separable convolutions were utilised to build StoneNet. The model achieved 97.98% accuracy. Furthermore, it increased efficiency by reducing complexity and computing cost [2]. In contrast, earlier methods that relied on traditional machine learning struggled to generalise across datasets. A comparison study introduced in 2021 evaluated deep convolutional neural networks (DCNNs) against traditional classifiers. DCNNs obtained better results, with a precision of 98% and a recall of 97% [3]. Coronal CT images used by a deep learning model for automated kidney stone detection attained an accuracy of 96.82%, emphasising the performance of CNNs with medical images [4].

To address inconsistencies among reported results, controlled comparisons across architectures are needed. A lightweight model for detecting renal stones could run in real time because of its efficiency. This model achieved a 96% F1 score while maintaining the false-negative rate of 4% [5]. However, most previous studies evaluate a single architecture or use different datasets, making direct comparison difficult. A researcher incorporated feature fusion to enhance the classification accuracy. As a result, using feature fusion techniques increased detection accuracy by 11% [6]. Another study evaluated data fusion techniques for the classification of kidney stones [7].

DNN was also implemented with multi-view imaging by using multiple angles of CT scan images. Consequently, this method increased the accuracy of kidney stone detection [8]. Moreover, neural networks were optimised for renal stone detection. As a result, they obtained a higher accuracy [9]. In 2022, researchers used deep learning with multiple CT image planes to detect kidney stones. This approach contributed to improving processing speed [10]. CNNs achieved better detection results than traditional methods based on varying metrics [11]. Researchers utilised deep learning with CT images to detect kidney stones. They used the XResNet50 architecture. It consists of ResLayer blocks, which decrease the resolution of the CT image by half. The accuracy of this detection model was 96.82%. In addition, image augmentation was used to avoid overfitting during the training stage [12].

In [13], researchers used two models. The first model was StackedEnsembleNet, which integrated four neural networks to increase detection accuracy. The second model was PSOWeightedAvgNet, which was responsible for finding the optimal weights for each neural network. Another model fusing ResNet101 and a custom CNN architecture. This model successfully classified stones, tumours, and normal tissues [14]. Guided Deep Metric Learning (GDML) enhanced the accuracy of stone detection. It reduced the need for large labelled datasets [15]. The transfer-based ensemble was proposed to improve classification accuracy for kidney diseases [16]. This approach was based on a combination of four models, including DarkNet19, InceptionV3, ResNet101, and YOLO. In [17], a CNN-LSTM hybrid model was utilised for ultrasound image interpretation. The model's accuracy for detecting kidney stones was 97%. Despite these advances, there is limited research directly comparing widely used architectures under identical conditions. Inception-ResNetV2 model was used for stone detection using kidney–ureter–bladder (KUB) images, while a ResNet and U-Net were used for accurately localising renal stones [19].

This study focuses strictly on binary CT image classification rather than object detection or localisation. Unlike detection-based approaches that aim to identify stone positions, the current work evaluates image-level classification performance under consistent experimental settings. The research compared three different deep learning architectures. These DNNs were chosen for their effective performance in medical applications. The comparison was performed using various evaluation metrics, including accuracy, precision, recall, and computational cost. As a result, it identified the DNN with the highest overall performance for kidney stone detection across all the performance metrics. Therefore, the main research gap addressed in this work is the lack of systematic comparison between commonly used architectures on multiple datasets using unified preprocessing and evaluation settings.

Material and methods

The comparison of deep learning architectures consists of several stages. The first stage is data collection. The data were divided into training and test datasets. The datasets were then used to train and test different DNNs.

A. Data Collection

In this research, two public datasets were utilized. The first dataset, obtained from [12], is referred to as Dataset 1, while the second, obtained from the Kaggle challenge[1], is called Dataset 2. The images stored in these datasets are CT (computed tomography) scans. Both datasets are publicly accessible; Dataset 1 is provided in [12], and Dataset 2 is available on Kaggle. All CT images are anonymised and contain no personal identifiers, ensuring compliance with patient privacy and dataset usage policies. CT scans are medical images acquired using a technique that produces cross-sectional images of the human body, as shown in Figure 1.

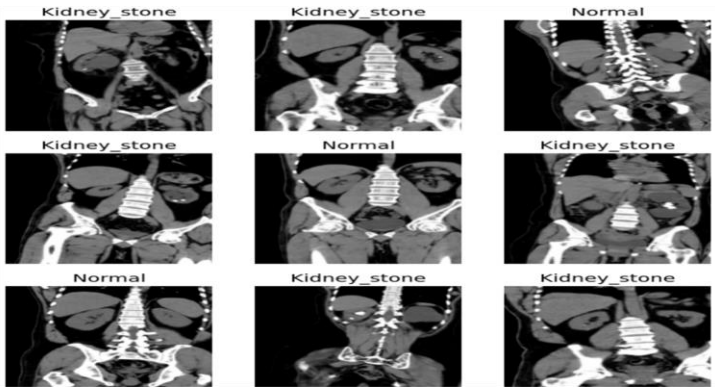

Figure 1: Example images in Dataset 1

B. Data Processing

In this study, two datasets were utilised. Each dataset was divided into three groups. Each dataset was randomly divided into 70% training and 30% testing sets at the image level. Due to the absence of patient identifiers in the publicly available datasets, splitting was performed at the image level. We acknowledge that patient-wise splitting would provide a more rigorous evaluation protocol. The data were batched using the DataBlock function with a 64-default batch size. The function is found in the fast.ai library [20], which prepares data for use in machine learning. Another component in fast.ai is called RandomResizedCropGPU, which is responsible for data augmentation. The algorithm implements augmentation by randomly resizing and cropping images during training, preserving the original aspect ratio. Data augmentation prevents the model from overfitting during training. All images were resized to 512×512 pixels in this study.

C. Model Training

Several experiments were performed to evaluate neural network architectures. Consequently, the best model for kidney stone detection from the three suggested architectures was highlighted.

a) ResNet50

ResNet50, proposed by Microsoft Research in 2015, is a 50-layer convolutional neural network. Its architecture includes residual connections, which enable the network to skip certain layers. Therefore, it accelerates the training process. These connections also prevent the model from experiencing the vanishing gradient problem. The model has been successfully implemented in various computer vision applications, such as object detection, semantic segmentation, and image classification. Table 1 presents the parameters of ResNet50.

Table 1: ResNet50 parameters

Total Parameters	25,557,032
Total Trainable Parameters	25,557,032
Total non-Trainable Parameters	0
Loss Function	Cross Entropy Loss
epochs	40

b) XResNet50

This neural network architecture was presented by Tong He et al. [21]. It has the same architecture as ResNet. Additionally, the neural network uses both 1×1 and 3×3 convolutions. It also has batch normalization before activation functions. As a result, these components improved the accuracy of object detection. XResNet50, with a deeper architecture, enables the network to learn more complex

[1] https://www.kaggle.com/datasets/mansoordaku/ckdisease

features. XResNet50 achieved state-of-the-art performance in several computer vision tasks. Table 2 shows the parameters of XResNet50.

Table 2: DenseNet201 parameters

Total Parameters	20,013,928
Total Trainable Parameters	20,013,928
Total Non-Trainable Parameters	0
Loss Function	Cross-Entropy Loss
epochs	40

c) DenseNet201

This convolutional neural network was presented by Gao Huang et al. [22]. The DenseNet family is known for its densely connected layers. Information flows directly through these layers. DenseNet201 contains 201 layers. The network architecture was built from a series of dense blocks. Each consists of several connected convolutional layers.Features are passed between consecutive dense blocks. The connection between dense blocks effectively enables feature reuse and avoids the vanishing gradient problem. The parameters of DenseNet201 are shown in Table 3. Training of the previous three models was performed in a cloud-based environment using Google Colab. This tool provides access to high-speed resources for code execution on the available GPUs (Graphical Processing Units). The trained models were saved for use in kidney-stone detection in the next stage.

Table 3: DenseNet201 parameters

Total Parameters	20,013,928
Total Trainable Parameters	20,013,928
Total Non-Trainable Parameters	0
Loss Function	Cross-Entropy Loss
epochs	40

Results and discussion

This section reports the evaluation of the three deep learning architectures on two datasets separately, using precision, recall, and F1 score metrics.

A. Results on Dataset 1

DenseNet201 demonstrated high performance according to the confusion matrix. The model correctly classified 196 examples as kidney stones, while 19 instances were wrongly classified as normal images. The model correctly classified 165 normal images and misclassified 16 images, as shown in Figure 2. The confusion matrix illustrates true versus predicted classes, allowing visual inspection of false positives and false negatives between models. For evaluation metrics, the detection accuracy on dataset 1 was 91%. The precision for normal images was 90%, and 92% for kidney stone images. The recall metric was 91% for both normal and kidney stone images. The F1 score was 92% for kidney stones and 90% for normal instances.

The ResNet50 model successfully classified 171 kidney stone images and misclassified 44. For normal images, the model identified 149 normal images and misclassified 32 as kidney stone images, as shown in Figure 3. The model achieved a detection accuracy of 81%. The precision was 77% for normal images and 84% for kidney stones. For the recall metrics, they were 82% for normal images and 80% for kidney stone images. The F1 score was 80% for normal images and 82% for kidney stone images.

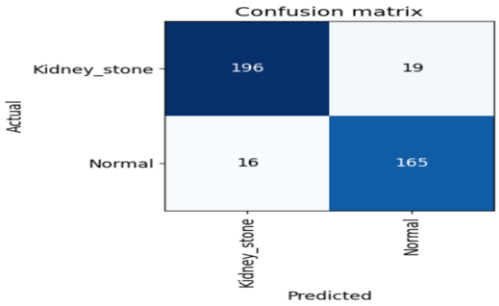

Figure 2: Confusion matrix of DenseNet201 on Dataset 1

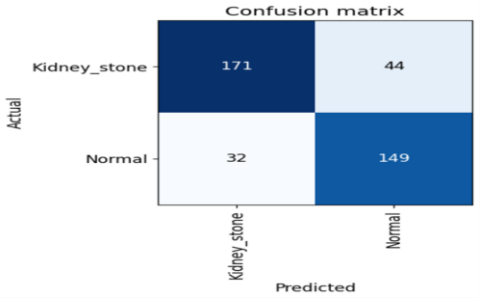

Figure 3: Confusion matrix of ResNet50 on Dataset 1

The XResNet50 had the highest performance among the three suggested network architectures. The model correctly classified 209 out of 215 kidney stone images, reflecting a low rate of false positives. It correctly classified 174 normal images and misclassified 7 images, as shown in Figure 4. The precision of this model is 97% for both normal and kidney stone images in the test dataset. The recall is 96% for normal and 97% for kidney stone images. The model achieved an accuracy of 97%. In addition, the F1 score was 96% and 97% for the normal and kidney stone images, respectively. In conclusion, the XResNet50 model outperformed other neural network architectures on Dataset 1. Since the reported accuracy values are relatively high, there remains a possibility of overfitting, particularly for XResNet50 on Dataset 1, despite using data augmentation and validation splitting to reduce this risk.

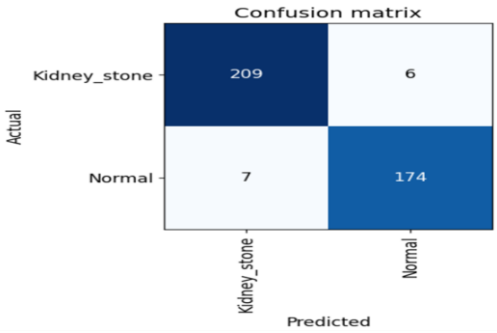

Figure 4: Confusion matrix XResNet50 on Dataset 1

B. Results on Dataset 2

When evaluating DenseNet201 on Dataset 2, it correctly recognised 153 kidney stone images and misclassified two images within normal classes. Therefore, the DenseNet201 performed better on Dataset 2 than on Dataset 1. In contrast, for normal images, the model correctly identified 86 images and incorrectly classified 36 normal images as containing kidney stones, as shown in Figure 5. The detection accuracy of DenseNet201 was 86%. The model reported a high precision of 98% for normal images 81% for kidney stone images. For the recall metric, it was 99% for kidney stone images and

only 70% for normal images. The F1 score was 89% for kidney stone images and 82% for normal images. The significant fluctuations in the metric values illustrate instability in the performance of the DenseNet201 model on Dataset 2.

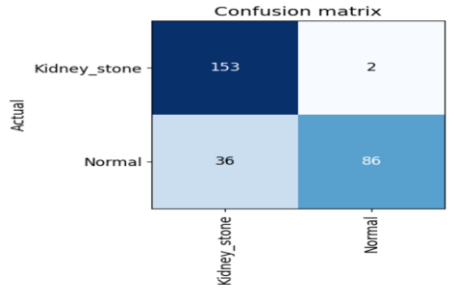

Figure 5: Confusion matrix of DenseNet201 on Dataset 2

TheResNet50 model correctly identified 152 kidney stone images and misclassified 3 as normal images. It also successfully predicted 92 images as belonging to the normal class, while incorrectly classifying 30 instances as kidney stone images, as shown in Figure 6. According to the evaluation metrics, the detection accuracy was 88%. The precision was 97% for normal images and 84% for kidney stone images. For the recall metric, the model achieved 98% for kidney stone images and 75% for normal images. The F1 score was 90% for kidney stone images and 85% for normal images.

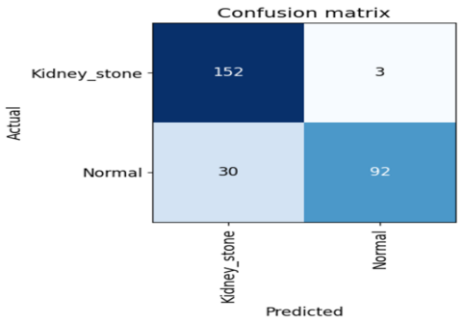

Figure 6: Confusion matrix of ResNet50 on Dataset 2

The XResNet50 architecture outperformed the two previous models on the datasets used in this research. The model correctly classified 153 kidney stone images and misclassified only 2 instances as normal images. For normal images, it correctly predicted 101 instances and misclassified 21 images, as illustrated in Figure 7. For evaluation metrics, XResNet50 achieved a detection accuracy of 92%. The precision metric was 88% for kidney stone images and 98% for normal images. Recall was 99% for kidney stone images and 88% for normal images. The model obtained an F1 score of 93% for kidney stone images and 90% for normal images. The XResNet50 demonstrated superior performance compared to other architectures. The summaries of performance comparisons for the three models are shown in Tables 4 and 5. Unlike prior studies that focus on a single architecture, this comparison provides practical guidance by revealing performance differences when models are evaluated under identical experimental conditions.

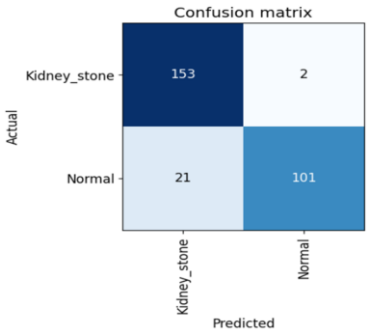

Figure 7: Confusion matrix of XResNet50 on Dataset 2

Table 4: Performance comparison on Dataset 1

Model	Accuracy	Precision (KS)	Recall (KS)	F1 (KS)	Precision (Normal)	Recall (Normal)	F1 (Normal)
ResNet50	0.81	0.84	0.80	0.82	0.77	0.82	0.80
DenseNet201	0.91	0.92	0.91	0.92	0.90	0.91	0.90
XResNet50	0.97	0.97	0.97	0.97	0.97	0.96	0.96
KS =Kidney Stone							

Table 5: Performance comparison on Dataset 2

Model	Accuracy	Precision (KS)	Recall (KS)	F1 (KS)	Precision (Normal)	Recall (Normal)	F1 (Normal)
ResNet50	0.88	0.84	0.98	0.90	0.97	0.75	0.85
DenseNet201	0.86	0.81	0.99	0.89	0.98	0.70	0.82
XResNet50	0.92	0.88	0.99	0.93	0.98	0.88	0.90
KS = Kidney Stone							

Extensive research has been conducted to compare the performance of deep learning architectures for the automatic detection of kidney stones. Yildirim et al. [12] compared various CNN architectures, including DenseNet and ResNet variants, on Dataset 1, which was also used in this research. Their results demonstrated the superior performance of ResNet50 over other models. Al-mutairi et al. [13] conducted a comparison study using fused deep-learning networks in kidney stone detection. They reported that the combination of features from multiple architectures contributed to enhancing the performance. Although they used a different dataset, the XResNet50 achieved better results compared to the standard DenseNet201 and ResNet50 architectures. Mehta and Wang [14] presented hybrid deep-learning frameworks to classify kidney stones. Their findings indicated that the combination of sequential models (ResNet101) and CNN-based feature extractors led to an enhancement in the performance of detection results. This study has several limitations. First, data splitting was performed at the image level rather than the patient level due to dataset constraints, which may introduce potential bias. Second, experiments were conducted using a single train-test split without repeated statistical validation. Third, external validation on independent multi-centre datasets was not performed. Future work will address these aspects to strengthen the robustness and generalizability of the findings.

Conclusion

This research investigated three deep-learning architectures used with medical images to find the most appropriate model for kidney stone detection. Therefore, it will help specialists during the diagnosis process. CT images were chosen for deep learning models due to their high accuracy. After comparing three deep learning models, the XRresNet50 achieved the highest performance among the evaluated architectures under the defined experimental setup. However, further validation is required to confirm generalizability across broader clinical settings. Two datasets were utilised for this study. Therefore, additional datasets are recommended for a more comprehensive comparison. Future work includes comparing techniques for kidney stone localisation. For example, GRAD-CAM (Gradient-weighted Class Activation Mapping) can locate regions of interest in medical images. This visual representation

might help doctors understand the logic behind decisions made by deep learning models. Across the two datasets, XResNet50 achieved the highest accuracy (97% and 92%), precision (97% and 98%), and F1-scores (97% and 93%), confirming its robustness relative to DenseNet201 and ResNet50.

References

[1] Ravisankar, P., Balaji, V., & Hameed, T. S. (2024). Deep learning-based renal stone detection: A comprehensive study and performance analysis. Applied Computer Systems, 29(1), 112–116. https://doi.org/10.2478/acss-2024-0014

[2] Yildirim, K., Bozdag, P. G., Talo, M., Yildirim, O., Karabatak, M., & Acharya, U. R. (2023). StoneNet: An efficient lightweight model based on depthwise separable convolutions for kidney stone detection from CT images. Frontiers in Genetics, 14. https://doi.org/10.3389/fgene.2023.1172456

[3] Lopez-Tiro, F., et al. (2021). Assessing deep learning methods for the identification of kidney stones in endoscopic images. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1234–1237). https://doi.org/10.1109/EMBC46164.2021.9630211

[4] Khan, M. A., Rauf, A., Fatima, B., & Khan, A. (2021). Deep learning model for automated kidney stone detection using coronal CT images. Computers in Biology and Medicine, 136, 104710. https://doi.org/10.1016/j.compbiomed.2021.104710

[5] Wang, F., Silvestre, G., & Curran, K. M. (2023). Lightweight framework for automated kidney stone detection using coronal CT images. arXiv. https://doi.org/10.48550/arXiv.2311.14488

[6] Villalvazo-Avila, E., Lopez-Tiro, F., & Daul, C. (2022). Improved kidney stone recognition through attention and multi-view feature fusion strategies. arXiv. https://doi.org/10.48550/arXiv.2211.02967

[7] Ochoa-Ruiz, G., Lopez-Tiro, F., & Daul, C. (2022). Comparing feature fusion strategies for deep learning-based kidney stone identification. arXiv. https://doi.org/10.48550/arXiv.2206.00069

[8] Roy, S., & Ahmed, A. (2022). Deep learning-based kidney stone detection using multi-view imaging. arXiv. https://doi.org/10.48550/arXiv.2205.01068

[9] Khan, A., & Miah, R. (2023). Optimizing neural networks for kidney stone detection on CT images. arXiv. https://doi.org/10.48550/arXiv.2301.06090

[10] Caglayan, A., Horsanali, M. O., Kocadurdu, K., Ismailoglu, E., & Guneyli, S. (2022). Deep learning model-assisted detection of kidney stones on computed tomography. International Brazilian Journal of Urology, 48(5), 830–839. https://doi.org/10.1590/S1677-5538.IBJU.2021.0845

[11] Danilovic, A., et al. (2021). Evaluation of convolutional neural networks for kidney stone detection. Journal of Endourology, 35(4), 567–573. https://doi.org/10.1089/end.2020.0895

[12] Yildirim, K., Bozdag, P. G., Talo, M., Yildirim, O., Karabatak, M., & Acharya, U. R. (2021). Deep learning model for automated kidney stone detection using coronal CT images. Computers in Biology and Medicine, 135, 104569. https://doi.org/10.1016/j.compbiomed.2021.104569

[13] Almutairi, M., Khan, S. H., & Zhang, L. (2024). An optimized fusion of deep learning models for kidney stone detection from CT images. Journal of King Saud University – Computer and Information Sciences, 36(2), 215–224. https://doi.org/10.1016/j.jksuci.2023.09.012

[14] Mehta, R. K., & Wang, P. S. (2025). Hybrid deep learning framework for classification of kidney CT images: Diagnosis of stones, cysts, and tumors. arXiv. https://doi.org/10.48550/arXiv.2502.04367

[15] Gupta, S., Li, F., & Becker, J. (2024). A metric learning approach for endoscopic kidney stone identification. Expert Systems with Applications, 234, 120834. https://doi.org/10.1016/j.eswa.2023.120834

[16] Kumar, N., Iqbal, M., & Ramachandran, V. (2025). A reliable kidney stone detection method using inductive transfer-based ensemble deep neural networks. The Bioscan, 20(1 Suppl.), S1-09–S1-16. https://doi.org/10.63001/tbs.2025.v20.i01.S1.pp09-16

[17] Chen, Y., Patel, A., & Singh, H. B. (2025). AI-powered early detection of kidney stones using a hybrid CNN-LSTM model. Journal of Neonatal Surgery, 14(2), 120–127.

[18] Verma, P., Zhao, H., & Yilmaz, M. (2023). Design and validation of a deep learning model for renal stone detection and segmentation on kidney–ureter–bladder images. Bioengineering, 11(1), 55. https://doi.org/10.3390/bioengineering11010055

[19] Howard, J., & Gugger, S. (2020). fastai: A layered API for deep learning. Information, 11(2), 108. https://doi.org/10.3390/info11020108

[20] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 558–567). https://doi.org/10.1109/CVPR.2019.00065

[21] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243