A Hybrid CNN Framework for Kidney Stone Detection Using Transfer Learning and Feature Fusion

Coşku Öksüz
Dept. of Electrical and Electronics Engineering
Izmir Bakircay University
Izmir, TURKEY
https://orcid.org/0000-0001-7116-2734

Artun Narter
Dept. of Electrical and Electronics Engineering
Izmir Bakircay University
Izmir, TURKEY
https://orcid.org/0009-0003-4275-6391

Bünyamin Ece
Dept. of Radiology
Kastamonu University
Kastamonu, TURKEY
https://orcid.org/0000-0001-6288-8410

Mustafa Koyun
Dept. of Radiology
Kastamonu University
Kastamonu, TURKEY
https://orcid.org/0000-0002-9811-4385

İsmail Taşkent
Dept. of Radiology
Kastamonu University
Kastamonu, TURKEY
https://orcid.org/0000-0001-6278-7863

M. Kemal Güllü
Dept. of Electrical and Electronics Engineering
Izmir Bakircay University
Izmir, TURKEY
https://orcid.org/0000-0003-2310-2985

***Abstract*— In this study, a deep learning method for kidney stone detection is proposed. The method utilizes transfer learning by extracting features from a pre-trained ImageNet model. However, unlike traditional transfer learning, which directly applies or fine-tunes a pre-trained model, the proposed approach integrates a custom-designed CNN that operates in parallel with the pre-trained network. The feature maps obtained from both networks are fused to enhance the model's representation power. After this integration, task-specific classification layers are added, and the training process is conducted on both the classification layers and the optimized model. This approach improves the overall performance of the model while providing a more efficient training process. As part of this study, a new dataset was created, consisting of 2166 axial slice images from 241 patients and 2018 axial slice images from 46 healthy individuals. Experiments conducted using EfficientNetV2s, MobileNetV4s, SqueezeNet, and ResNet18-based models revealed that the EfficientNetV2s and MobileNetV4s-based models excelled in terms of accuracy, while the SqueezeNet and ResNet18-based models provided stronger results in terms of interpretability.**

Keywords—Kidney stone, detection, classification, deep learning, transfer learning

I. Introduction

Kidney stones are solid structures formed by the accumulation of minerals and salts in the kidneys, which can cause severe pain and various health issues as they move through the urinary tract [1], [2]. The disease stemming from kidney stones, i.e., nephrolithiasis, is a common health problem worldwide, with research indicating that approximately 10-12% of the adult population is affected [3]. Particularly in regions with hot and arid climates, where inadequate fluid intake and poor dietary habits prevail, the prevalence may be even higher [4]. Kidney stones are 2-3 times more common in men compared to women [5]. In addition, one study stated that individuals between the ages of 40-70 are at risk [1].

The early diagnosis of kidney stones plays a crucial role in patient health outcomes. Stones detected in the early stages are generally smaller in size, making them more likely to be treated through medication, fluid intake, or minimal invasive methods. In contrast, undetected stones can grow and cause obstructions in the urinary tract, leading to severe pain, urinary tract infections, and kidney function deterioration. Long-term undiagnosed kidney stones can lead to increased pressure in the kidneys, resulting in hydronephrosis (swelling of the kidney) and, in more advanced cases, irreversible kidney failure. Furthermore, approximately 50% of individuals with kidney stones are at risk of recurrence within 5-10 years [6]. Therefore, early diagnosis allows for the implementation of lifestyle and dietary changes to prevent stone formation, thus reducing the risk of recurrence. In cases diagnosed early, stones can be treated naturally or through less invasive methods without the need for surgery. Consequently, individuals with suspected kidney stones should consult a specialist without delay and ensure regular health check-ups.

Accurate and reliable diagnostic methods are essential for the detection of kidney stones. To this end, physicians commonly use ultrasound and computed tomography (CT) scans. While ultrasound has the advantage of not involving radiation, it has limitations in precisely determining the size, location, and number of stones. In this context, CT imaging stands out due to its higher accuracy rates. Computed tomography (CT) is an effective method for detecting and evaluating kidney stones, providing clear information about the size, location, and number of stones. Particularly, non-contrast CT (NCCT) can detect even non-calcium stones, allowing for the identification of all types of stones [7]. However, the major drawback of CT scans is the radiation involved, which should be minimized by avoiding unnecessary repetition.

In recent years, artificial intelligence (AI) and machine learning-based approaches have gained prominence in medical imaging. In the diagnosis of kidney stones, machine learning overcomes the limitations of traditional methods by offering faster and more accurate analyses. AI-assisted systems reduce dependence on radiologist interpretations, decrease error rates, and can identify the chemical

This work is supported by the Scientific and Technological Research Council of Turkey. Project number: 123E442

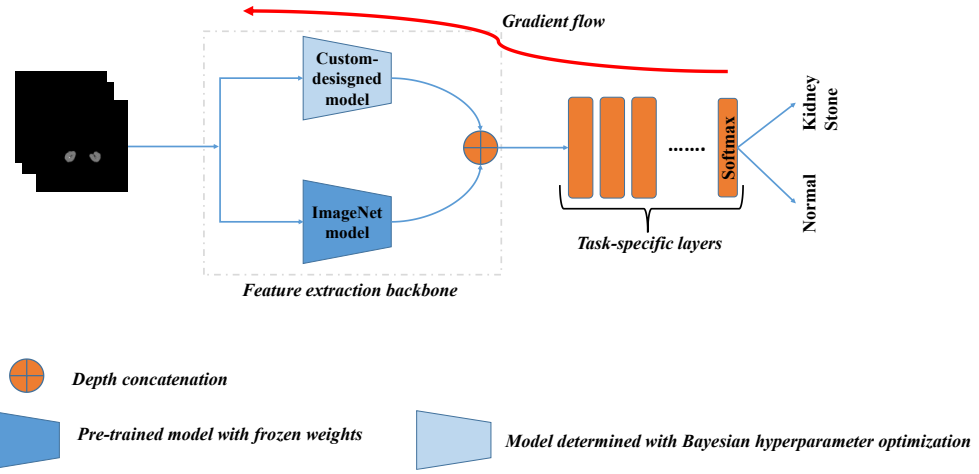

Fig. 1. The proposed method. The method utilizes a custom-designed network, determined through Bayesian hyperparameter optimization, alongside a pre-trained model with completely frozen weights. It accepts the ROI image (i.e., the kidneys) for processing.

composition of stones, enabling personalized treatment plans. Moreover, these technologies help prevent unnecessary surgical interventions, enhance the efficiency of healthcare services, and improve patient quality of life.

In [8], a total of 790 NCCT images were collected from 278 patients diagnosed with kidney stones, while 1009 NCTT images were obtained from 165 healthy individuals. All images were acquired in the coronal plane. In this study, the XResNet50 model, developed by modifying ResNet50 from the residual network family, achieved accuracy, sensitivity, precision, and F1 scores of 96.82%, 97%, 97%, and 97%, respectively, without utilizing a transfer learning strategy. The study conducted in [9], developed a hybrid method called *ExDark19* for kidney stone detection in which deep learning- and machine learning-based methods have been utilized. The model utilizes feature extraction based on DarkNet19. The most informative features were selected using Iterative Neighbourhood Component Analysis (INCA) and classified with the k-Nearest Neighbors (kNN) algorithm. The study achieved an accuracy of 99.22% using 10-fold cross-validation and 99.71% with hold-out validation on the data set introduced in [8]. In [10], various transfer learning models, including MobileNet, Inceptionv3, InceptionResNetv2, and Xception, were considered for kidney stone detection. A weighted combination of the model predictions was used, with the particle swarm optimization method employed to determine the optimal weights. The accuracy, sensitivity, and F1 scores obtained on the dataset from [8] were 98.84%, 98.79%, and 98.79%, respectively. In reference work [11], a combined method was proposed, utilizing AlexNet as a feature extractor and Extreme Learning Machine (ELM) as a classifier. The weight optimization for ELM was performed using a modified firefly swarm optimization algorithm. The method achieved sensitivity, specificity, and F1 scores of 91.90%, 97.08%, and 99.72%, respectively.

The summarized literature indicates that most studies utilize transfer learning models and metaheuristic algorithms to optimize weight updates for improved classification performance. However, while most of these studies rely on existing models, there remains a need for novel and optimized approaches. In this study, a deep learning-based method is proposed for kidney stone detection. Unlike conventional transfer learning approaches, the model integrates a pretrained ImageNet-based network with a custom-designed CNN, whose architecture is optimized through hyperparameter tuning, operating in parallel. Feature maps from both networks are fused to enhance representation capability, followed by task-specific classification layers. The proposed method was evaluated on a large NCCT data set consisting of 4184 images. Experimental results demonstrate that the proposed method significantly outperforms its transfer learning counterparts.

The study is organized as follows: Section II presents the proposed method, Section III covers the experimental analysis, Section IV discusses the findings, and finally, the conclusion is provided in Section V.

II. Proposed Method

As seen in Fig. 1, the proposed method for kidney stone detection incorporates elements of transfer learning by leveraging a pre-trained ImageNet model for feature extraction. However, unlike traditional transfer learning, which typically fine-tunes or directly applies a pre-trained model, our approach integrates a custom-designed CNN, optimized through hyperparameter tuning, that operates in parallel with the pre-trained network. The feature maps from both networks are fused to enhance the representation power of the model. Following this integration, task-specific classification layers are added in the following order: batch normalization, ReLU activation, dropout, adaptive average pooling, and two dense layers. The first dense layer consists of 90 feature channels, while the second one outputs two classes for classification. The training process is conducted only on the task-specific layers and the hyperparameter-optimized model, ensuring both generalization and task-specific adaptability. This means, as depicted with the red line in Figure 1., gradient computations are performed only for the task-specific layers and the optimized model, allowing for efficient training while preserving the learned features of the pre-trained network.

A. Custom designed model

In this study, a custom model is designed to learn task-related patterns. A network diagram is given in Fig. 2, explaining how the custom model is formed in the study. As seen in Fig. 2, it begins with a core block comprising two sequential repetitions of 2D convolution, batch normalization, ReLU, and pooling layers. During the optimization process,

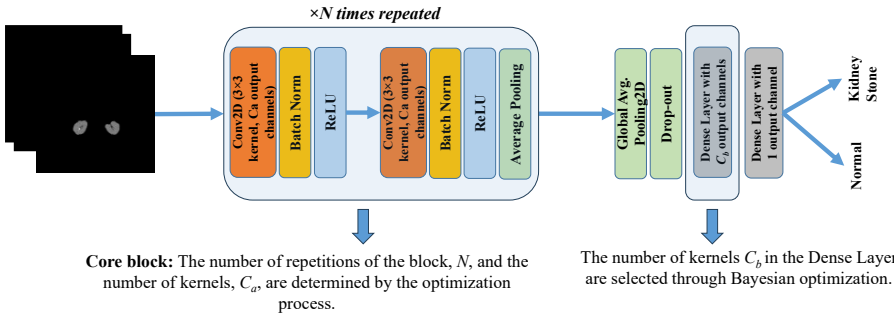

Fig. 2. The network diagram that explains the model creation process within the study.

this core block is repeated a selected number of times within the search range. In determining the number of filters in the 2D convolution layer, the elements of the set C_a={32, 64, 96, 128, 160, 192, 224, 256} are considered. Finally, the model is completed by incorporating global average pooling, dropout, and two dense layers. While the number of neurons C_b in the first dense layer is searched within the range of 30 to 150, it is set to one for the binary classification problem.

B. *Pre-trained model*

Pre-trained models have undergone extensive training to recognize complex feature patterns. To leverage this prior knowledge, the feature maps of a custom-designed model are enriched by integrating those from a pre-trained model. These pre-trained feature maps contribute high-level details, ultimately enhancing the model's ability to capture meaningful patterns. The pre-trained model is completely incorporated into the framework with frozen weights.

C. *Combining the custom-designed model with pretrained model*

At this stage, feature maps with a 7×7 spatial resolution are extracted from both the custom-designed and pre-trained networks. Therefore, any layers producing smaller feature maps are removed from each branch. Depth-wise feature map concatenation is adopted instead of alternative fusion strategies such as element-wise summation or attention mechanisms, as it preserves complementary features from both networks without introducing additional learnable parameters. After concatenation, additional layers are appended to enable the network to learn task-specific high-level representations from the fused tensor, enriched by the distinct characteristics of each branch. This fusion strategy differs from conventional hybrid methods by integrating structurally and functionally distinct paths—one optimized for domain-specific texture learning, the other for capturing generic deep features. Such structured heterogeneity facilitates richer and more robust representations with minimal computational overhead, contributing to improved performance and generalization.

III. Experimental Analysis

A. *Data set*

The experiments were conducted on a private data set, which is currently not publicly available, retrospectively collected from Kastamonu Research and Training Hospital in Turkey. This data set consists of 4,184 NCCT images, including 2,166 from abdominal scans of 241 patients and 2,018 from 46 healthy individuals, all acquired in the axial plane. Radiologists defined kidney masks for each slice, as well. Since only the region of interest, i.e., the kidneys were considered in our study, the kidney region was preserved by masking each CT abdominal image in the data set.

To ensure unbiased model evaluation, the data set was divided into six non-overlapping folds at the patient level. One-fold was allocated as the hold-out set, while the remaining five were used for five-fold cross-validation (CV5). The distribution of patient and control subjects, along with the corresponding number of images in each fold, is summarized in Table I. This setup guarantees that no overlap exists between training, validation, and test subjects, ensuring subject-level independence across folds.

TABLE I. Fold-wise Distribution of Subjects and Images

Fold	Patients	Controls
Fold #1	283 Images/39 Patients	302 Images/7 Healthy
Fold #2	311 Images/40 Patients	374 Images/8 Healthy
Fold #3	385 Images/39 Patients	367 Images/8 Healthy
Fold #4	362 Images/39 Patients	340 Images/8 Healthy
Fold #5	385 Images/39 Patients	343 Images/8 Healthy
Hold-out	440 Images/45 Patients	292 Images/7 Healthy

B. *Performance metrics*

The metrics used to evaluate the performance are provided between Equations (1) and (6).

$$Recall = TP/TP + FN \quad (1)$$

$$Accuracy = TP + TN/TP + TN + FP + FN \quad (2)$$

$$Specificity = TN/TN + FP \quad (3)$$

$$Precision = TN/TP + FP \quad (4)$$

$$F_1 = 2Recall.Precision/Recall + Precision \quad (5)$$

$$\kappa = P_{observed} - P_{chance}/1 - P_{chance} \quad (6)$$

C. *The hyperparameter optimized model*

The custom model is determined using the Bayesian optimization method, with a total of 30 objective function evaluations conducted during the experiments. Each model, created based on the network diagram given in Fig. 2, was trained using the Adam optimizer for 50 epochs. The learning rate and mini-batch size was set to 10−4 and 32, respectively.

TABLE II. COMPARISON OF THE PRE-TRAINED MODELS AND THE PROPOSED MODELS BASED ON PERFORMANCE METRICS.

Models	Accuracy (%)		Recall (%)		Kappa (κ) (%)		Specificity (%)		Precision (%)		F1 (%)	
	CV5	Holdout	CV5	Holdout	CV5	Holdout	CV5	Holdout	CV5	Holdout	CV5	Holdout
MobileNetv4s	88,7	91	93,3	92,8	77,3	81,2	84,1	88,4	85,9	92,4	88,7	90,6
SqueezeNet	92,7	94,7	95,0	97,1	85,4	88,9	90,5	91,1	91,0	94,3	92,7	94,1
ResNet18	93,5	94,6	95,7	96,4	87,0	88,6	91,3	91,8	91,9	94,7	93,5	94,1
EfficientNetV2S	94,4	94,9	95,9	96,9	88,6	89,2	92,7	91,8	93,2	94,7	94,3	94,3
Proposed (SqueezeNet–based)	95,8	97,6	96,8	**98,2**	91,5	**94,9**	94,7	96,6	94,8	97,8	95,8	97,4
Proposed (ResNet18–based)	95,8	**97,9**	96,9	**98,2**	91,5	95,5	94,7	97,3	94,9	98,2	95,8	97,8
Proposed (MobileNetv4s–based)	96,5	97,3	95,9	95,9	93,0	94,3	**97.0**	**99,3**	**97,1**	**99,5**	96,5	**97,7**
Proposed (EfficientNetv2s–based)	**97,0**	97,6	**97,0**	97,5	**93,9**	**94,9**	96,9	97,7	96,9	98,4	**97,0**	97,6

A randomly selected 10% of the training set was reserved as a validation set to follow up the model training process. At the end of the optimization process, the final architecture consists of a core block repeated six times, resulting in 12 convolutional layers with 128, 32, 96, 64, 256, and 224 filters, respectively. Additionally, the first dense layer was determined to have 90 neurons.

D. The performance analysis of the proposed framework

The pre-trained models, i.e., MobileNetv4s [12], EfficientNetV2s [13], SqueezeNet [14], and ResNet18 [15], were considered in this study. Accordingly, four different versions of the proposed framework were developed using these lightweight pre-trained models. Each version was trained for 50 epochs using the Adam optimizer with a global learning rate of 10−4. For baseline models, the final classification layer of each pre-trained network was replaced with a dense layer suitable for binary classification. During training, only this final layer was updated with a learning rate 10× higher than the global rate, while the remaining layers were fine-tuned using the global learning rate. In Table II, the average CV5 scores based on each performance metric is given comparatively with the transfer learning models of the pre-trained networks. The performance scores on the hold-out are provided, as well. The experimental results in Table II indicate that the proposed method consistently outperforms its baseline counterparts across all evaluation metrics, in both CV5 and holdout. In terms of cross-validation (CV5) performance, the proposed models show considerable improvements over the baseline pre-trained networks. For example, the EfficientNetV2S-based model achieves a CV5 accuracy of 97.0%, which is an improvement of +2.6% compared to the baseline EfficientNetV2S model at 94.4%. Similarly, its recall increases from 95.9% to 97.0% (+1.1%), and the κ value improves from 88.6% to 93.9% (+5.3%). Specificity also increases from 92.7% to 96.9% (+4.2%), and precision from 93.2% to 96.9% (+3.7%). The ResNet18-based proposed model demonstrates similar gains in CV5 performance, with an accuracy increase from 93.5% to 95.8% (+2.3%). The κ value rises from 87.0% to 91.5% (+4.5%), and specificity improves from 91.3% to 94.7% (+3.4%). Precision also shows an increase from 91.9% to 94.9% (+3.0%), confirming the robustness of the proposed improvements in cross-validation testing. For the SqueezeNet-based version, CV5 accuracy rises from 92.7% to 95.8% (+3.1%), with notable improvements in recall (95.0% to 96.8%, +1.8%), κ (85.4% to 91.5%, +6.1%), specificity (90.5% to 94.7%, +4.2%), and precision (91.0% to 94.8%, +3.8%). The MobileNetv4s-based model achieves the highest relative improvement in CV5, with accuracy increasing from 88.7% to 96.5% (+7.8%). Its κ value improves significantly, from 77.3% to 93.0% (+15.7%), and specificity sees an impressive gain from 84.1% to 97.0% (+12.9%). Precision also increases from 85.9% to 97.1% (+11.2%). As it is seen in Table II, in the hold-out testing, similar performance trends are observed, confirming the generalizability of the proposed framework.

IV. DISCUSSION

The experimental results in Table II emphasize the significant performance gains achieved by the varying versions of the proposed method. In Fig. 3., the gradient-weighted class activation maps (GradCAM) are given, as well, for each version of the proposed method. As seen in the GradCAM visualizations for the SqueezeNet-based model in Fig. 3(a), the model focuses more on the kidney containing the stone. Similarly, the GradCAM visualization for the ResNet18-based model in Fig. 3(b) shows stronger activation around the pathological region, although some activations are also observed near the healthy kidney. On the other hand, the GradCAM visualizations for the EfficientNetV2S-based model in Fig. 3(c) and the MobileNetV4-based model in Fig. 3(d) indicate that these models fail to focus on the relevant regions.

Based on the results presented in Table II, the EfficientNetV2S- and MobileNetV4-based models appear to be the most effective versions of the proposed method across all metrics. However, their GradCAM visualizations, shown in Fig. 3(c) and Fig. 3(d), do not align with these findings, suggesting a discrepancy between quantitative performance and localization quality. More specifically, the MobileNetV4-based model (Fig. 3d) exhibits a broader distribution of attention rather than precisely localizing the affected kidney. On the other hand, the EfficientNetV2S-based model (Fig. 3c) fails to focus on the region of interest, indicating that its high performance may stem from global feature utilization rather than localized pathological indicators. This behaviour contrasts with their higher recall and accuracy values, raising concerns about the interpretability of their decision-making process. The SqueezeNet-based model (Fig. 3a) and the ResNet18-based model (Fig. 3b) demonstrate a strong focus on the kidney with stones, indicating their ability to accurately highlight pathological regions. This aligns with their high recall scores (96.8% and 96.9%, respectively), suggesting that

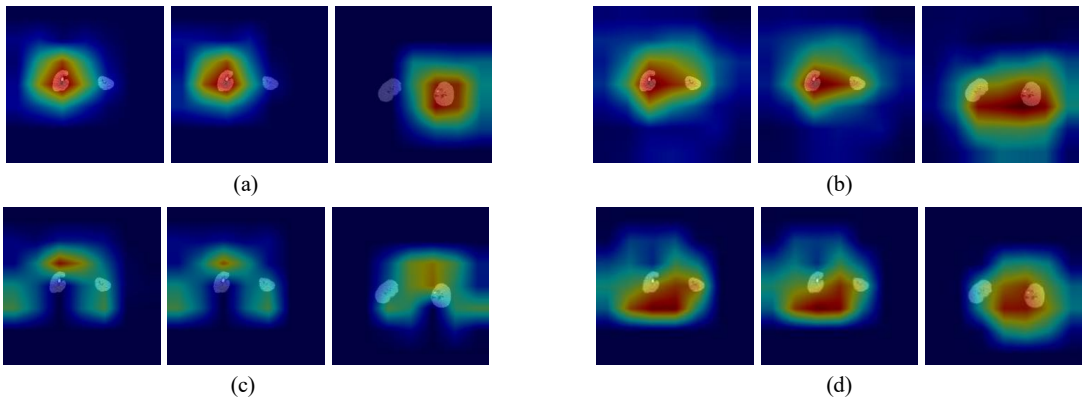

Fig. 3. The gradient-weighted class activation maps of the proposed models based on different architectures. (a) SqueezeNet-based model. (b) ResNet18-based model. (c) EfficientNetV2s-based model. (d) MobileNetV4-based model.

these models effectively capture positive cases. These results indicate a clear trade-off: the SqueezeNet- and ResNet18-based models prioritize sensitivity, making them more suitable for detecting pathological cases, whereas the EfficientNetV2s- and MobileNetV4s-based models are not suitable candidates as the pre-trained model of the proposed method.

V. CONCLUSION

In this study, we developed and evaluated a hybrid CNN framework for kidney stone detection using transfer learning and feature fusion. The proposed method integrates a custom-designed CNN with pre-trained lightweight models, leveraging their feature extraction capabilities to enhance classification performance on a private CT abdomen data set. The results showed that the proposed method, particularly when leveraging EfficientNetV2S and MobileNetV4s, significantly outperformed baseline models in all evaluation metrics, such as accuracy, recall, and κ. However, despite the quantitative improvements, the GradCAM visualizations indicated that these models often failed to localize the region of interest, suggesting a potential gap between classification performance and interpretability. In contrast, the SqueezeNet and ResNet18-based models, although exhibiting lower overall performance, excelled in focusing on the pathological regions, which may make them more suitable for applications requiring high sensitivity and interpretability. Overall, while EfficientNetV2S- and MobileNetV4s-based models achieved the best classification results, their limited interpretability poses challenges in clinical use where decision transparency is vital. Additionally, the single-center data set restricts generalizability due to the lack of external validation. Future work will focus on developing a ROI-based model to enhance both accuracy and clinical interpretability.

ACKNOWLEDGMENT

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Grant No. 123E442.

REFERENCES

[1] L. Moftakhar, F. Jafari, M. Ghoddusi Johari, R. Rezaeianzadeh, S. V. Hosseini, and A. Rezaianzadeh, 'Prevalence and risk factors of kidney stone disease in population aged 40–70 years old in Kharameh cohort study: a cross-sectional population-based study in southern Iran', BMC Urol., vol. 22, no. 1, p. 205, Dec. 2022, doi: 10.1186/s12894-022-01161-x.

[2] S. Nikpay, K. Moradi, M. Azami, M. Babashahi, M. Otaghi, and M. Borji, 'Frequency of Kidney Stone Different Compositions in Patients Referred to a Lithotripsy Center in Ilam, West of Iran', J. Pediatr. Nephrol., vol. 4, no. 3, Art. no. 3, Dec. 2016, doi: 10.22037/jpn.v4i3.13121.

[3] T. Alelign and B. Petros, 'Kidney Stone Disease: An Update on Current Concepts', Adv. Urol., vol. 2018, p. 3068365, Feb. 2018, doi: 10.1155/2018/3068365.

[4] R. Siener, 'Nutrition and Kidney Stone Disease', Nutrients, vol. 13, no. 6, p. 1917, Jun. 2021, doi: 10.3390/nu13061917.

[5] S. R. Khan et al., 'Kidney stones', Nat. Rev. Dis. Primer, vol. 2, p. 16008, Feb. 2016, doi: 10.1038/nrdp.2016.8.

[6] S. Shastri, J. Patel, K. K. Sambandam, and E. D. Lederer, 'Kidney Stone Pathophysiology, Evaluation and Management: Core Curriculum 2023', Am. J. Kidney Dis., vol. 82, no. 5, pp. 617–634, Nov. 2023, doi: 10.1053/j.ajkd.2023.03.017.

[7] W. Brisbane, M. R. Bailey, and M. D. Sorensen, 'An overview of kidney stone imaging techniques', Nat. Rev. Urol., vol. 13, no. 11, pp. 654–662, Nov. 2016, doi: 10.1038/nrurol.2016.154.

[8] K. Yildirim, P. G. Bozdag, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, 'Deep learning model for automated kidney stone detection using coronal CT images', Comput. Biol. Med., vol. 135, p. 104569, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104569.

[9] M. Baygin, O. Yaman, P. D. Barua, S. Dogan, T. Tuncer, and U. R. Acharya, 'Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images', Artif. Intell. Med., vol. 127, p. 102274, May 2022, doi: 10.1016/j.artmed.2022.102274.

[10] S. Asif, X. Zheng, and Y. Zhu, 'An optimized fusion of deep learning models for kidney stone detection from CT images', J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 7, p. 102130, Sep. 2024, doi: 10.1016/j.jksuci.2024.102130.

[11] H. Ding, Q. Huang, and N. Razmjooy, 'An improved version of firebug swarm optimization algorithm for optimizing Alex/ELM network kidney stone detection', Biomed. Signal Process. Control, vol. 99, p. 106898, Jan. 2025, doi: 10.1016/j.bspc.2024.106898.

[12] D. Qin et al., 'MobileNetV4 -- Universal Models for the Mobile Ecosystem', Sep. 29, 2024, arXiv: arXiv:2404.10518. doi: 10.48550/arXiv.2404.10518.

[13] M. Tan and Q. V. Le, 'EfficientNetV2: Smaller Models and Faster Training', Jun. 23, 2021, arXiv: arXiv:2104.00298. doi: 10.48550/arXiv.2104.00298.

[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size', Nov. 04, 2016, arXiv: arXiv:1602.07360. doi: 10.48550/arXiv.1602.07360.

[15] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. Accessed: Jan. 31, 2025. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html