Computers in Biology and Medicine 135 (2021) 104569

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep learning model for automated kidney stone detection using coronal CT images

Kadir Yildirim [a], Pinar Gundogan Bozdag [b], Muhammed Talo [c], Ozal Yildirim [c,*], Murat Karabatak [c], U.Rajendra Acharya [d,e,f]

[a] Department of Urology, Faculty of Medicine, University of Turgut Ozal, Malatya, Turkey
[b] Department of Radiology, Elazig Fethi Sekin City Hospital, Elazig, Turkey
[c] Department of Software Engineering, Firat University, Elazig, Turkey
[d] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore
[e] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan
[f] School of Management and Enterprise University of Southern Queensland, Springfield, Australia

ARTICLE INFO

Keywords:
Kidney stone
Medical image
Deep learning
Computed tomography

ABSTRACT

Kidney stones are a common complaint worldwide, causing many people to admit to emergency rooms with severe pain. Various imaging techniques are used for the diagnosis of kidney stone disease. Specialists are needed for the interpretation and full diagnosis of these images. Computer-aided diagnosis systems are the practical approaches that can be used as auxiliary tools to assist the clinicians in their diagnosis. In this study, an automated detection of kidney stone (having stone/not) using coronal computed tomography (CT) images is proposed with deep learning (DL) technique which has recently made significant progress in the field of artificial intelligence. A total of 1799 images were used by taking different cross-sectional CT images for each person. Our developed automated model showed an accuracy of 96.82% using CT images in detecting the kidney stones. We have observed that our model is able to detect accurately the kidney stones of even small size. Our developed DL model yielded superior results with a larger dataset of 433 subjects and is ready for clinical application. This study shows that recently popular DL methods can be employed to address other challenging problems in urology.

1. Introduction

Kidney stone disease is one of the most common health problems, although the frequency varies between different countries. In prevalence studies, this rate is reported to be between 1 and 20% [1,2]. Kidney stones can lead to kidney failure, loss of workforce by causing severe pain and decrease in the quality of life by obstructing the urinary system [3,4]. For example, more than 2 million people in the United States of America (USA) apply to the emergency department every year for renal colic or stone-related back pain. Approximately half of these patients undergo non-contrast computed tomography (NCCT) [5]. The studies on the diagnosis of this disease can improve the quality of life of patients and possible kidney failure [6].

Although the selection of right imaging technique for the detection of kidney stones varies according to the clinical situation and patient-related parameters, however, it is still the first step in the diagnosis of the disease [7]. Ultrasonography (USG), kidney ureter bladder (KUB), NCCT are used to detect urinary system disease. NCCT has become the standard imaging modality for the diagnosis of acute flank pain [1]. In a meta-analysis, low-dose NCCT was found to detect with a sensitivity and specificity of 93.1% and 96.6%, respectively [8].

Nowadays, deep learning (DL) techniques have been successfully applied to various fields using medical images and physiological signals. The deep models have been used successfully employed in many areas such as segmentation of medical images [9,10], classification [11–14], and lesion detection [15,16]. Various types of medical images like magnetic resonance imaging (MRI), computed tomography (CT) and X-ray have been used to develop accurate and robust DL models to aid the clinicians in their diagnosis of diseases such as Covid-19, cardiac arrhythmia, prostate cancer, brain tumor, skin, and breast cancer [9–19]. DL techniques are also employed in the urology field for automatic detection of ureteral stones and kidney stones. Fitri et al. [20]

* Corresponding author.
E-mail address: ozal@firat.edu.tr (O. Yildirim).

https://doi.org/10.1016/j.compbiomed.2021.104569
Received 20 April 2021; Received in revised form 1 June 2021; Accepted 9 June 2021
Available online 14 June 2021

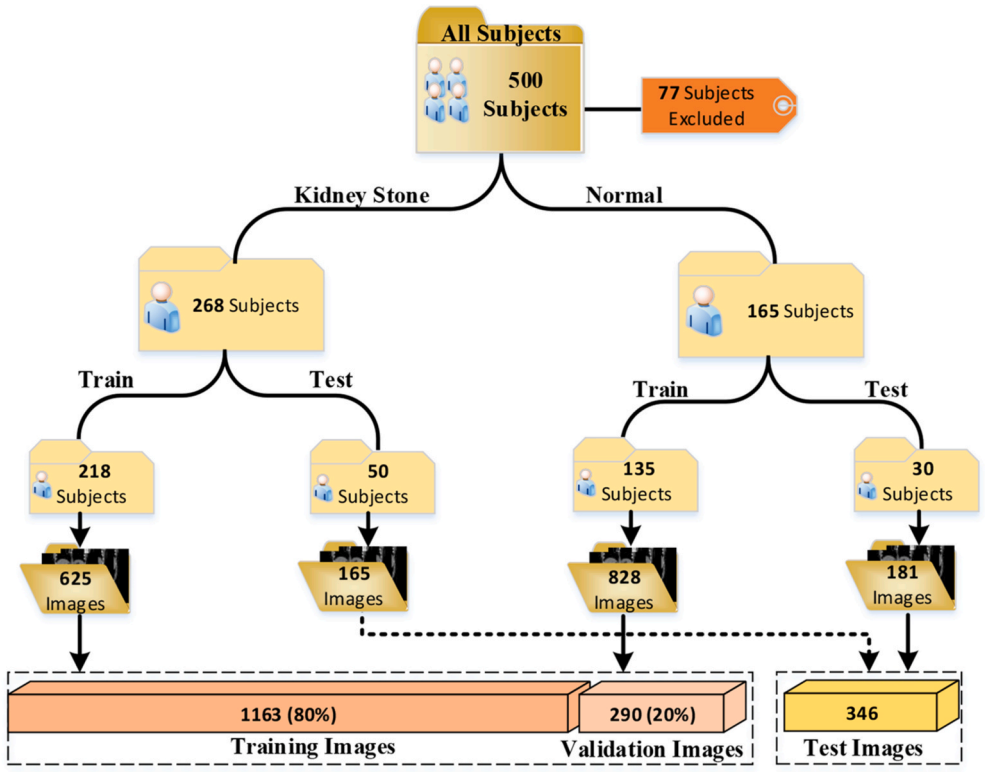

Fig. 1. Data distributions used for training and testing in this study.

proposed a CNN-based model to classify urinary stones from micro-computed tomography (micro-CT). They used a total of 2430 micro-CT slices and reported a test accuracy of 0.9959. Jendeberg et al. [21] developed a CNN model for classifying pelvic calcifications. They concluded that the CNN model results are better than the mean assessment by seven radiologists. Längkvist et al. [22] used a CNN model for identifying ureteral stones in thin slice CT volumes.

In this study, a model was designed to prevent the missed stone diagnosis using CT images and minimize the physician-induced errors by using DL techniques. Most of the stone patients will be admitted to the emergency room and it may be difficult to get a specialist radiologist always. Sometimes the reporting period of computed tomography can be

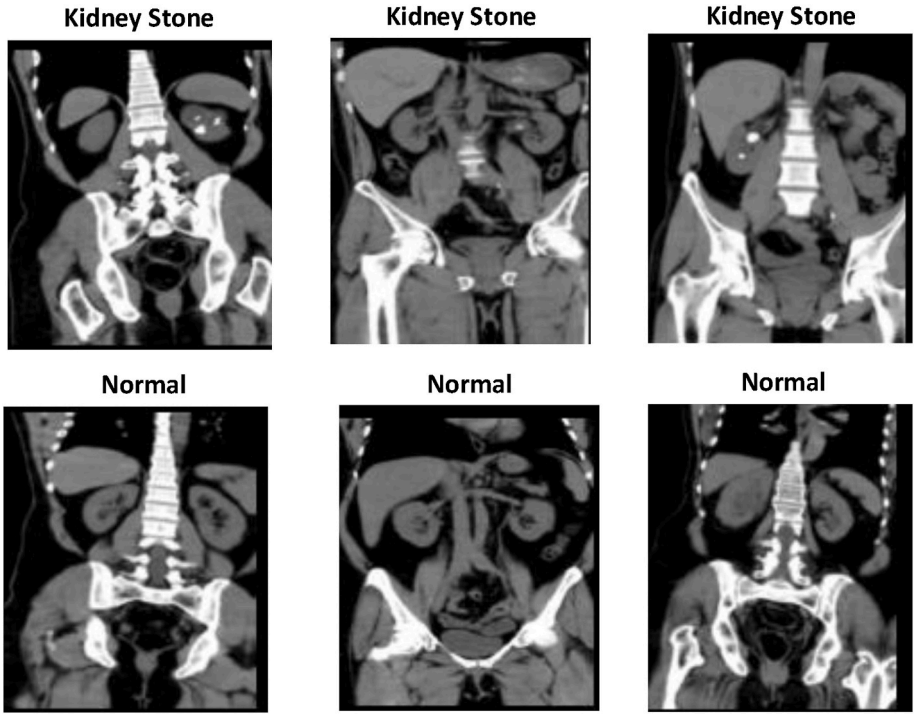

Fig. 2. Typical examples of normal and kidney stone CT images obtained using various augmentation techniques.

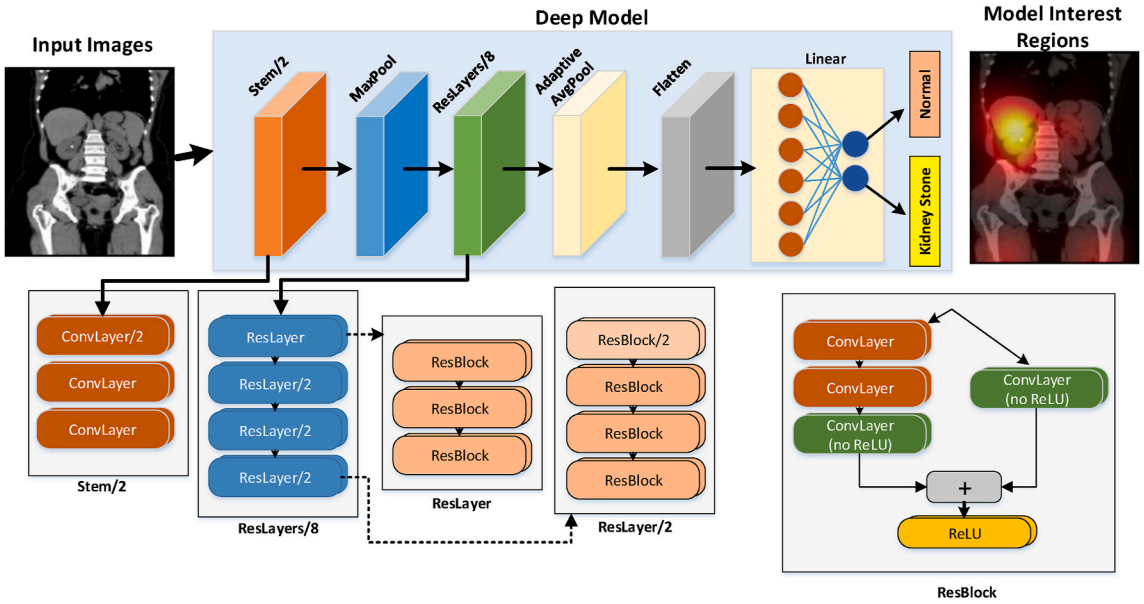

Fig. 3. Block representation showing the layer organization of XResNet-50 deep learning model employed for automated kidney stone detection.

delayed due to the scarcity of radiologists. Misinterpretation may be encountered in cases where imaging is performed faster than usual [23]. In addition, since computed tomography reporting is a time-consuming process, the evaluation of kidney with computer-aided methods for early and accurate diagnosis will be a significant contribution to the health field.

Urology is one of the branches of medicine which is using advanced technology for accurate diagnosis. The robots and endourological interventions have been employed in surgery in the field of urology [24]. Nowadays, DL have been exceedingly used in healthcare application due to sudden explosion of data. In this study, we have made an effort to implement DL for the automated classification of kidney stone cases using CT images. The model can detect kidney stones. Also the proposed model can focus on kidney stone area on CT images. Therefore, the model has two different tasks: (i) detection and (ii) localization.

2. Patients and methods

2.1. Data

This study was conducted after obtaining approval from the ethics committee of Fırat University, Turkey. We have collected 500 NCCT images taken from the patients admitted for urinary system stone disease to Elazığ Fethi Sekin City Hospital, Turkey. All images were acquired in the supine position using a single scanner Philips Healthcare, Ingenuity Elite (Netherland), without contrast administration. The dataset and code are available at https://github.com/yildirimozal/Kidney_stone_detection.

Personal information of the patients were not used and coronal CT sections (CT protocol 120 kV; auto tube current 100 mA–200 mA; 128 mm detector; mAs 203; slice thickness 5 mm) were evaluated separately by a radiologist and urologist. The experts (radiologist and urologist) carried out the labeling process by specifying whether there are stones without performing any segmentation on the CT images. Patients aged between 18 and 80 years were included in this study. Sixty-seven patients with double-j (pigtail) ureteral catheters, under 18 and over 80 years of age, having single kidney, kidneys with anomalies, and atrophic kidneys were excluded from this study. In this work, patients with kidney stones are considered as patient group and subjects without kidney stones as control group. A total of 433 subjects, 278 stone positive and 165 normal, were used in this study. The CT images obtained from different sections of these patients were used in the study. Seven hundred ninety images for patients diagnosed with kidney stone and 1009 images of normal subjects were obtained. Subjects used in the training and validation phase were not used for the testing phase to prevent biased results. In other words, the subjects used in the test and train stages are completely different. The CT images used for training, validation and testing used in this study are shown in Fig. 1. Typical examples of normal and kidney stone CT images obtained using various augmentation techniques are shown in Fig. 2.

2.2. Deep model

The DL is the sub-branch of artificial intelligence and has recently shown remarkable achievements in various areas. In this study, we used the cross-residual network (XResNet-50) model for kidney stone detection [25]. The training of XResNet-50 deep model was carried out on raw CT images from scratch. We used augmentation techniques such as rotation (10°) and zooming on the raw images to prevent the model to memorize input images of same patient during training (overfitting problem). Hence, during model training, various augmented images were randomly fed to the model in each epoch. XResnet-50 architecture consists of four stages. The image resolution is reduced to half in Stem and MaxPooling layers. In addition, each stage has ResLayer blocks which reduces the resolution by half, and these blocks consist of multiple layers. The layer block and the model architecture of XResnet model used in this study are shown in Fig. 3. The model is fed with raw CT images as input, provides the output class in the output layer and also the region of interest on which the model has concentrated to obtain accurate diagnosis. The training of XResnet-50 model was carried out on the Google Colab tool using Fastai (v2) library developed on Pytorch deep learning library [26]. The Adam optimization algorithm [27] and cross-entropy loss were used to adjust the parameters of XResNet-50 model.

3. Results

During the training phase of the model, 80% of 1453 C T images were used for training and remaining 20% for validation. After the completion of model training, test performance values were obtained using 346 images that were not used during the training of the deep model. A graph of accuracy rate and loss values for each epoch during training of

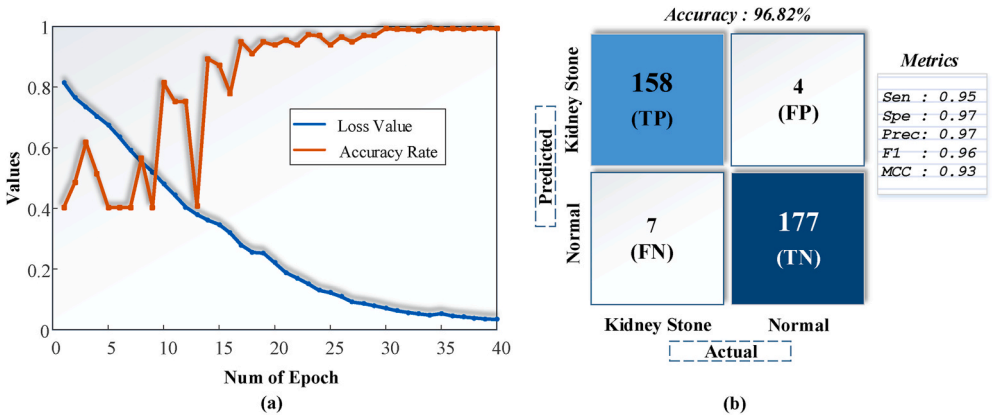

Fig. 4. (A) Graphs of loss values and accuracy rates obtained for various epochs during the training of model, and (b) confusion matrix obtained on the test data.

Table 1
Summary of performance parameters obtained using our proposed DL model.

	Precision	Recall	F1-Score	Support
Kidney Stone	0.98	0.96	0.97	165
Normal	0.96	0.98	0.97	181
Accuracy			0.97	346
Macro Avg	0.97	0.97	0.97	346
Weighted Avg	0.97	0.97	0.97	346

DL model is shown in Fig. 4 (a). The model continued learning on training data for 40 epochs. The confusion matrix obtained using the test data is given in Fig. 4 (b).

It can be seen from the confusion matrix that, the model predicted 158 kidney stone images correctly (true positive, TP) and 7 kidney stone images in normal class (false negative, FN). Also, the model correctly classified 177 images (true negative, TN) as normal class. On the other hand, it has misclassified 4 normal images (False Positive, FP) to kidney stone class. The performances used to evaluate the model are precision (TP/(TP + FP)), recall (TP/(TP + FN)), F1-score (2 × precision × recall/(precision + recall)) and accuracy ((TP + TN)/(FP + FN)). Evaluating these metrics together is an important criterion in measuring the

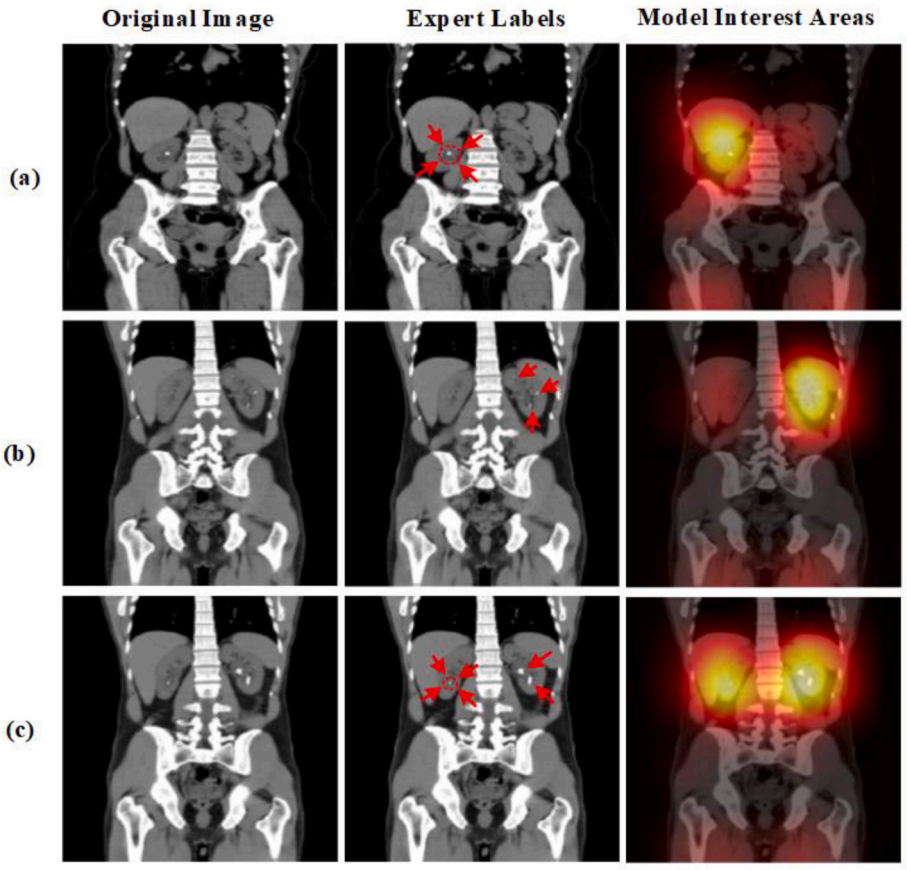

Fig. 5. Sample test images showing the areas on which the DL model has concentrated for diagnosis. Red arrows were regions used by experts to show the stones in the images.

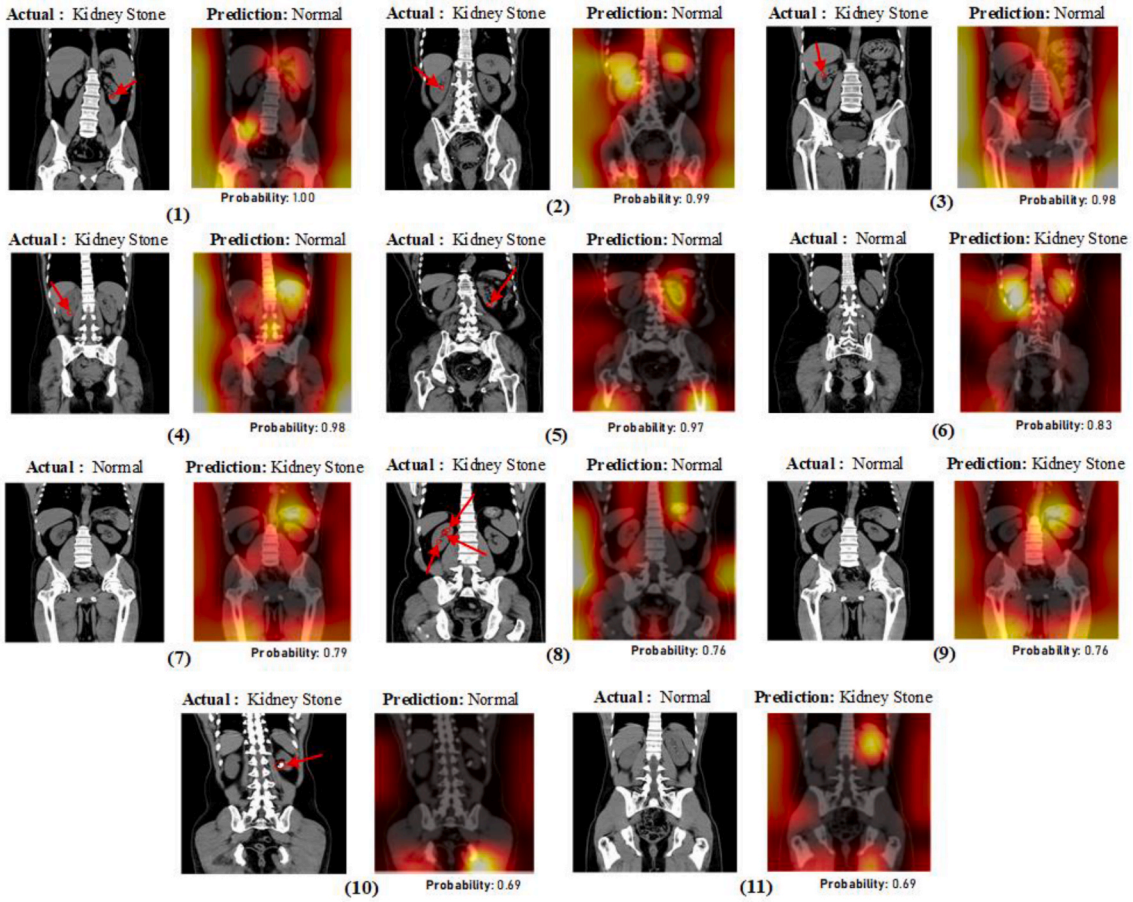

Fig. 6. Images misclassified by the model, the CT images fed as input to the model (left), the interested regions selected by the model for classification (right), and the stones marked on the CT image by the experts (red arrows).

performance of the model. While the precision indicates the correct prediction rate of kidney stone cases, recall indicates how accurately the cases that need to be predicted as positive are predicted correctly. F1-Score gives the harmonic average of precision and recall values. The model yielded 96.82% accuracy rate and 95.76% sensitivity using 146 test cases. The summary of performance metrics obtained using our proposed model using the testing data is given in Table 1.

Although the accuracy values obtained by DL models are the important measurement performance indicators, it is important to know the kind of criteria the deep model has made due to the black-box structures. Thus, it is possible to create more reliable models and employ them in healthcare applications. Recently, intensive research has been carried out to determine the areas where deep models have been used. In this study, we used Grad-CAM [28] application to determine the areas on which our model concentrated to obtain highest classification performance.

It can be noted from Fig. 5 that, few images presented show the areas (region of interest) the DL model concentrated on to classify it to "Kidney Stone" class. Red arrows were used to show the stones in the images to compare the model output with expert opinion. In Fig. 6, *eleven* images were misclassified by the model during diagnosis and the regions concentrated on the images by the model.

4. Discussions

In this study, we used coronal NCCT sections which includes the whole abdomen, pelvis, part of the thorax, and lower extremity. The kidney tissue in the area of interest constitutes a small part of the total area. Despite this, the proposed DL model detected accurately cases with small size of kidney stones, with a sensitivity and specificity of 95% and 97%, respectively.

The rib tip closely related to the lower pole of kidney entering the cross-sectional area may have caused the model to give erroneous results and hence wrongly predicted as stones by the model as shown in Fig. 6. It can be noted from the heat-map of the model that the model has concentrated on lower pole of kidney and rib. It can be noted from Fig. 6 that, in images numbered 7 and 9, the model focused on the stomach part which is close to the upper pole of the left kidney. Millimetric opacities which created the impression of stone in the stomach may have caused the model to give erroneous results.

The image 11 of Fig- 6 is classified as erroneous by the model as it focused on the left kidney in the heat map. It can be noticed from these images that, there are millimetric calcified areas in the kidney. This section, which was not evaluated as a stone by the experts, may have been evaluated as positive by the model. We think that the results of these erroneous evaluations can be improved when more such sections are used to train the model. Among 165 images in the stone group, only seven images were detected as normal, although there were stones. When the six images (1, 2, 3, 4, 5, and 8) in Fig. 6 are examined by experts, it is seen that the existing stones can be hardly seen.

The image number 10 in Fig. 6, is misclassified even though there is a large stone. Other sections of the same patient were examined by the DL model, the model correctly detected all stones in the rest of images. In this misclassified section, it appears that the stone is outside the kidney. This may have caused the wrong diagnosis by our model.

There are many more organs in coronal sections compared to axial sections. In these sections, the area of kidney and stone within it constitutes much less total area. In addition, since the area to be scanned is

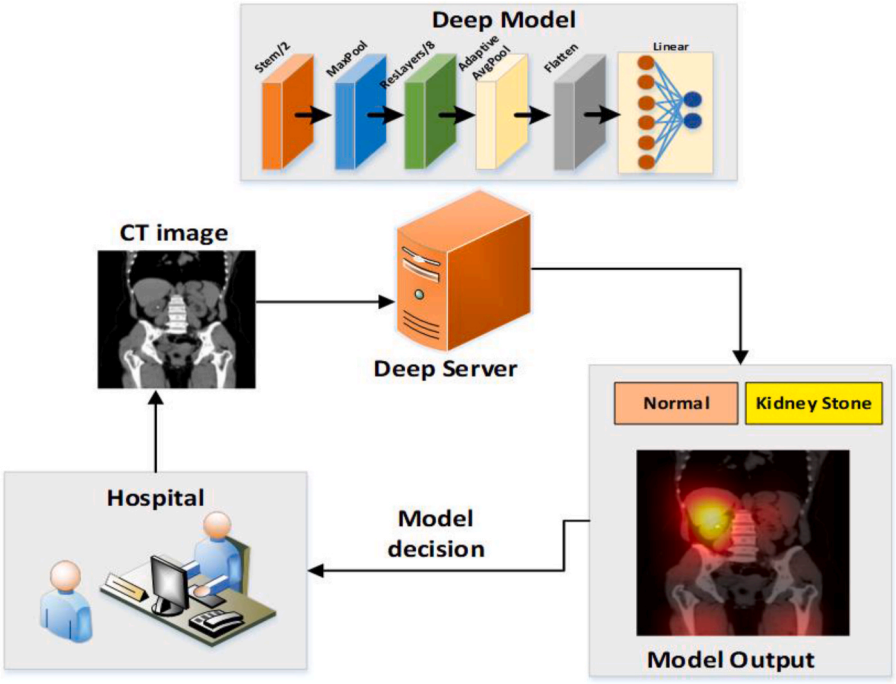

Fig. 7. Snapshot of the demonstration of deep learning model as an auxiliary tool in healthcare centers.

extensive, there are many variables in this area, such as external genitalia, bone structures, bladder, internal organs, and lungs. Nevertheless, the model focuses on the correct localization of almost all cases.

Evgeniya et al. showed that when radiologists report faster at peak pace, they make more statistically significant errors. It has been shown that 10% of the radiologists missed significant findings when they took approximately 10 min to screen an abdominal CT image and this error increased to 26% when the screening time is reduced to 5 min [23]. In general, it is known that approximately 4% of errors are made in clinical radiology practice [29]. Considering that more than 1 billion radiological evaluations are made in the world annually, it can be concluded that about 40 million wrong evaluations may be made wrong in a year [30,31]. Hence, each image is evaluated by two independent radiologists to reduce the error rate significantly [32]. However, it may be difficult to implement this due to shortage of radiologists and cost [30]. Hence, such accurate DL models can be used to replace one of the radiologists and obtain the second opinion.

The proposed model is clinically reliable diagnostic tool as it has obtained highest classification performance and also able to identify correct regions of interest on the image for accurate decision making. The model marks the areas with stones on CT images without the need for any segmentation process. An important advantage of the model in this study is that it provides an automatic detection and visual information about the localization of stones. The CT images obtained from different sections of these patients were used in the study. The model can classify cases according to the input images. These images can be fed as partial sections of the patients. We trained our model with a different section of the patients. Therefore, the model has the ability to detect different sections. The snapshot of the proposed deep model which can be used as an adjunct tool by the clinicians in healthcare centers is shown in Fig. 7.

The major limitation of this study is that all the images used in this study were collected from only one hospital which can limit the generalizability of the model. Hence, in future, we intend to collect images from different sources. We used CT images as it can be easy obtained from emergency centers. In future, we intend to evaluate our method using axial and sagittal planes of images. Also, the kidney stones can be classified according to the sizes. We intend to collect CT images with labeled stone sizes. This future work can provide information about relation between size and classification performance.

5. Conclusion

In this study, a DL model is proposed for the detection of kidney stone cases using CT images. The proposed deep model has yielded 96.82% accuracy rate using 433 subject's data. In addition to the classification using coronal CT images, our deep model has successfully marked the areas of interest during the decision-making process. The results were evaluated by two experts (one radiologist and one urologist). Clinically, the regions identified by the model was in agreement by our medical experts for most of the images. Hence, our proposed DL model is accurate and can assist the radiologists to detect kidney stone cases accurately using our developed model.

Disclosure statement

No competing financial interests exist.

Declaration of competing interest

There is no conflict of interest in this work.

References

[1] C. Türk, A. Petřík, K. Sarica, C. Seitz, A. Skolarikos, M. Straub, et al., EAU guidelines on diagnosis and conservative management of urolithiasis, Eur. Urol. 69 (2016) 468–474.
[2] A. Chewcharat, G. Curhan, Trends in the Prevalence of Kidney Stones in the United States from 2007 to 2016, Urolithiasis, 2020.
[3] K.L. Penniston, S.Y. Nakada, Development of an instrument to assess the health related quality of life of kidney stone formers, J. Urol. 189 (2013) 921–930.
[4] F. New, B.K. Somani, A complete world literature review of quality of life (QOL) in patients with kidney stone disease (KSD), Curr. Urol. Rep. 17 (2016) 88.
[5] E.S. Hyams, F.K. Korley, J.C. Pham, B.R. Matlaga, Trends in imaging use during the emergency department evaluation of flank pain, J. Urol. 186 (2011) 2270–2274.
[6] B.R. Matlaga, Toward a better understanding of kidney stone disease: platinum priorities, Cités (2012) 166–167.
[7] W. Brisbane, M.R. Bailey, M.D. Sorensen, An overview of kidney stone imaging techniques, Nat. Rev. Urol. 13 (2016) 654.

[8] H. Xiang, M. Chan, V. Brown, Y.R. Huo, L. Chan, L. Ridley, Systematic review and meta-analysis of the diagnostic accuracy of low-dose computed tomography of the kidneys, ureters and bladder for urolithiasis, J Med Imaging Radiat Oncol 61 (2017) 582–590.
[9] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imag. 32 (4) (2019) 582–596.
[10] H.R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, K. Mori, Deep learning and its application to medical image segmentation, Med. imaging Technol. 36 (2) (2018) 63–71.
[11] M. Talo, U.B. Baloglu, Ö. Yıldırım, U.R. Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, Cognit. Syst. Res. 54 (2019) 176–188.
[12] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Comput. Biol. Med. (2020) 103792.
[13] O. Kott, D. Linsley, A. Amin, A. Karagounis, C. Jeffers, D. Golijanin, et al., Development of a deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate cancer biopsies: a pilot study, Eur. Urol. Focus 7 (2) (2019) 347–351.
[14] E. Shkolyar, X. Jia, T.C. Chang, D. Trivedi, K.E. Mach, M.Q.-H. Meng, et al., Augmented bladder tumor detection using deep learning, Eur. Urol. 76 (2019) 714–718.
[15] Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers, DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning, J. Med. Imag. 5 (3) (2018), 036501.
[16] R. Kijowski, F. Liu, F. Caliva, V. Pedoia, Deep learning for lesion detection, progression, and prediction of musculoskeletal disease, J. Magn. Reson. Imag. 52 (6) (2020) 1607–1619.
[17] O. Yildirim, M. Talo, E.J. Ciaccio, R. San Tan, U.R. Acharya, Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records, Comput. Methods Progr. Biomed. 197 (2020) 105740.
[18] Y. Li, L. Shen, Skin lesion analysis towards melanoma detection using deep learning network, Sensors 18 (2018) 556.
[19] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, U.R. Acharya, Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images, Pattern Recogn. Lett. 133 (2020) 232–239.
[20] L.A. Fitri, F. Haryanto, H. Arimura, C. YunHao, K. Ninomiya, R. Nakano, M. Haekal, Y. Warty, U. Fauzi, Automated classification of urinary stones based on microcomputed tomography images using convolutional neural network, Phys. Med. 78 (2020) 201–208.
[21] J. Jendeberg, P. Thunberg, M. Lidén, Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network, Urolithiasis (2020) 1–9.
[22] M. Längkvist, J. Jendeberg, P. Thunberg, A. Loutfi, M. Lidén, Computer aided detection of ureteral stones in thin slice computed tomography volumes using Convolutional Neural Networks, Comput. Biol. Med. 97 (2018) 153–160.
[23] E. Sokolovskaya, T. Shinde, R.B. Ruchman, A.J. Kwak, S. Lu, Y.K. Shariff, et al., The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: a pilot study, J. Am. Coll. Radiol. 12 (2015) 683–688.
[24] P.F. Müller, D. Schlager, S. Hein, C. Bach, A. Miernik, D.S. Schoeb, Robotic stone surgery–current state and future prospects: a systematic review, Arab journal of urology 16 (3) (2018) 357–364.
[25] Jou B, Chang S-F. Deep cross residual learning for multitask visual recognition. Proceedings of the 24th ACM international conference on Multimedia2016. p. 998-1007.
[26] J. Howard, Gugger S. Fastai, A layered API for deep learning, Information 11 (2020) 108.
[27] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[28] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision2017. p. 618-626.
[29] A. Sabih, Q. Sabih, A.N. Khan, Image perception and interpretation of abnormalities; can we believe our eyes? Can we do something about it? Insights into imaging 2 (2011) 47–55.
[30] S. Waite, J. Scott, B. Gale, T. Fuchs, S. Kolla, D. Reede, Interpretive error in radiology, Am. J. Roentgenol. 208 (2017) 739–749.
[31] M.A. Bruno, E.A. Walker, H.H. Abujudeh, Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction, Radiographics 35 (2015) 1668–1676.
[32] L.H. Garland, On the scientific evaluation of diagnostic procedures: presidential address thirty-fourth annual meeting of the Radiological Society of North America, Radiology 52 (1949) 309–328.