2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU) | 978-1-6654-7723-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/WiDS-PSU57071.2023.00037

Classification of kidney abnormalities using deep learning with explainable AI

Khadiime Jhumka
Faculty of Information, Communication and Digital Technologies
University of Mauritius
Reduit, Mauritius
khadhiime@gmail.com

Maleika Heenaye-Mamode Khan
Faculty of Information, Communication and Digital Technologies
University of Mauritius
Reduit, Mauritius
m.mamodekhan@uom.ac.mu

Zahra Mungloo-Dilmohamud
Faculty of Information, Communication and Digital Technologies
University of Mauritius
Reduit, Mauritius
z.mungloo@uom.ac.mu

Swalay Aboo Fedally
Department of Nephrology
Sir Seewoosagur Ramgoolam National (SSRN) Hospital
Pamplemousses, Mauritius
drswalay@yahoo.com

***Abstract*—Kidney is a critical organ in the human body, and any damage to it causes substantial injury to the body. Early detection is essential. As a result, Artificial Intelligence, particularly deep learning, is being used to detect kidney disorders at an early stage. This paper focuses on the classification of kidney diseases using a deep learning model notably XResNet50. The dataset includes 9527 CT pictures in both the axial and coronal planes, and it is split into three categories: normal, stone, and tumor. The model was used to distinguish between these three classes using training and validation data, and it was applied to predict unseen testing data. The model achieved a testing accuracy of 97%, a precision of 0.97, a sensitivity of 0.97 and a F1-score of 0.97. Furthermore, the model was able to recognize the kidney region in the CT-scan of the Stone, Normal, and most tumor classes based on the SHAP images.**

Keywords—kidney stone, kidney tumor, deep learning, XResNet, explainable AI, SHAP

I. INTRODUCTION

Kidneys are essential organs for the correct functioning of the human body. They filter blood, eliminate waste via urine, and maintain the body's fluid levels. A kidney illness will bring serious harm to the body. There are many kidney diseases such as chronic kidney disease, kidney tumor and kidney stones to name a few. Chronic kidney disease affects more than 10% worldwide [1], was the 16th greatest cause of death in 2016 and is predicted to move up to the fifth spot by 2040 [2]. Kidney tumor is a severe renal condition that affects around 400,000 people each year [3]. According to the Global Cancer Observatory (GCO), this illness is responsible for more than 131,000 fatalities [4]. Meanwhile kidney stone disease is one of the most common health problems which can lead to kidney failure. During the period of 2017 to 2018, the prevalence of kidney stones was 10.9% in males and 9.5% in women [5]. Furthermore, with global warming, these figures will rise dramatically as has been shown by some recent studies [6–8]. Despite recent breakthroughs, patients with end-stage renal disease die at a greater rate than those who do not have the disease. According to a global case study, death rates range from 20% to 50% within 24 months [9]. To decrease future kidney instances, it is necessary to detect these disorders.

Artificial intelligence (AI) has shown to provide substantial improvements in health care [10,11]. Recent advancements in big data technologies, and artificial intelligence have made it easier to create strategies for automating medical activities. Kidney disease detection is a multidisciplinary field being investigated by IT professionals in order to aid medical staff by delivering trustworthy diagnostic findings with the help of deep learning [12–14]. Recently, given the large data storage capacities, there has been an expansion of image datasets and this has resulted in automated decision support systems using machine learning and deep learning. This, together with advancements in renal imaging methods, has resulted in research into the automated early identification and care for people with kidney abnormalities. An example is the study done by Sudharson et al. [15] which demonstrated the use of deep learning to classify kidney ultrasound images based on various kidney disorders such as kidney cysts, kidney stones, and kidney malignancies. However, AI has yet to establish its promising future in the medical field since doctors perceive it as a black box because they do not know what features the trained models use to categorize the medical images. Several attempts to handle this issue have been made including Explainable AI. Consequently, this work presents a deep learning model, based on the XResNet50 architecture, for classifying various kidney illnesses from CT scans, as well as the features that the trained model uses to distinguish between kidney abnormalities and a normal kidney.

The body of this study is structured as follows. Section II looks at the study that has been done on this topic. Section III covers the suggested solution and its constituents. Section IV focuses on the analysis of experimental results.

II. RELATED WORKS

With the advancement of current technology, several cutting-edge AI algorithms for the identification of kidney stones and kidney malignancies have emerged. The research done in this field is described briefly in the paragraphs that follow.

In a study, Zhou et al. [16] have used deep learning to identify between benign and malignant renal tumors in a research regarding detecting renal cancers based on deep learning. To conduct the classification of 192 CT scans, the

978-1-6654-7723-9/23/$31.00 ©2023 IEEE
DOI 10.1109/WiDS-PSU57071.2023.00037

convolutional neural network (CNN) architecture employed was pretrained InceptionV3, and the model's performance was evaluated using the receiver operating characteristic parameter. The model achieved a high level of accuracy of 97%. Pedersen et al. [17] developed a deep learning algorithm to classify kidney tumors in a study with 20,000 CT-labelled images from 369 individuals to detect oncocytoma. The dataset was randomly divided into training, validation, and testing at the patient level to prevent data leakage. The CNN architecture utilized was a fine tuned model of the ResNet50V2. The test findings demonstrated a high degree of efficiency, with scores of 97.3%, 93.5%, and 93.3% for accuracy, area under the curve, and specificity respectively.

On the other hand, Yildirim et al.[18] have developed a deep neural model for detecting kidney stones. Deep Neural Network (DNN) was used to detect abnormalities from computed tomography (CT) images and an accuracy of 96.82% and a sensitivity of 95.76% were achieved on 146 test cases. It was reported that the biggest weakness of this study was that all images were acquired from a single hospital and thus, could possibly limit the generalisability of the deep learning model. Such a system could potentially be more reliable if images were taken from axial and coronal planes as well, which is not the case.

da Cruz et al. [19] have come up with an automated precise segmentation solution of the kidney images to eliminate the variability between various specialists. The authors have used deep convolutional neural networks (CNN) on tomography images. The pre-trained network, U-Net, which is a feed forward network was applied on the images. This technique works on a symmetrical expansion path, which could locate the kidneys accurately. U-Net was fine-tuned and could segment the kidneys more efficiently compared to benchmarked research, leading to less false positives. Pre-trained models, which have already learned to extract powerful and informative features from a vast number of images of big datasets, are useful in the medical field due to the unavailability of large labelled datasets. In a work conducted by Sudharson and Kokil [20], the pre-trained ResNet-101 was adopted to extract the necessary kidney features from ultrasound images, after the removal of speckling noises. Support Vector Machine (SVM) was then employed as a classifier to accomplish the multi-class classification into normal, cyst, stone, and tumor. Ensemble learning has also been explored in the development of automated kidney problem detection applications. Sudharson and Kokil [21] have adopted the ensemble multiple-support vector machine (MSVM) classification model to categorise the kidney images. Data augmentation was applied to reduce overfitting. In this work, an accuracy of 89.53% was achieved.

The researchers concluded that deep learning algorithms can effectively classify kidney disorders and advocated the adoption of these diagnostic models to avoid overtreatment. However, despite the fact that deep learning produces excellent results, medical practitioners are hesitant to use them since there is no transparency in which factors contribute to the abnormality in question. As a result, explainable AI is evolving to close the explainability gap.

III. MATERIALS AND METHODS

Figure 1 depicts the proposed system solution. Two publicly available labelled CT scan datasets were utilised in this study. The procedure is as follows: (1) merging the two datasets (2) image pre-processing (3) training the classification model and hyperparameter tuning (4) analysing the results and (5) look at the features used for classification.

Fig. 1. Workflow of proposed solution

A. Datasets

In this work, two image datasets were used. The first one is collected from different hospitals in Bangladesh and has both coronal and axial CT scans and it is available on Kaggle (https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone) [22]. The second dataset is available on github (https://github.com/yildirimozal/Kidney_stone_detection) and it is collected from Turkey [18]. The datasets were classified into 3 classes: normal (5,077 images), kidney stone (2167 images) and kidney tumor (2283 images). The CT scans have both axial and coronal views as shown in Figure 2. Then the merged dataset was augmented through rotation and zooming. It was later preprocessed to the size of 512x512. Afterwards, it was randomly divided into 3 datasets: the training dataset, the validation dataset and the testing dataset. The first two sets of data are used to train the model and finetune the hyperparameters of the model and the testing dataset is used to assess the performance of the trained model.

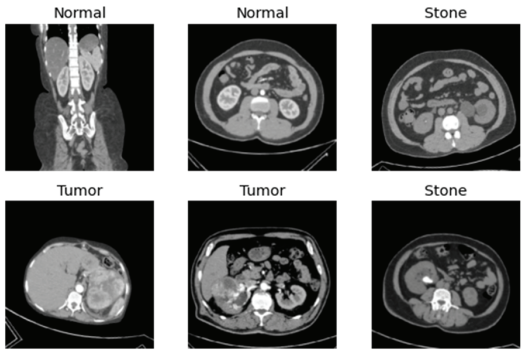

Fig. 2. Images from the two datasets

B. Classification Model

The model used for classification is the pretrained XResNet50 model from Fastai library. The model contains 50 convolution layers which are used for feature extraction

and the activation function used in this model is the ReLu function. The residual block consists of a pair of the combination of a convolution layer followed by a batch normalization layer and finished with a ReLU activation function and there are 14 residual blocks in the pre-trained model.

The input size to the model was 64 x 3 x 224 x 224 with the last layer giving the output size of 64 x 3, with 3 being the number of classes and 64 being the default batch size for the Fastai pre-trained model. The number of trainable parameters used was 25,635,168. Table 1 shows the most effective hyperparameters for the XResNet50 model. To avoid model overfitting, an early stopping function was introduced. After that, the model was trained using the training and validation data.

TABLE 1. HYPERPARAMETERS USED

Hyperparameters	Value
Learning rate	1×10^{-2}
Number of epochs	20
Loss function	Cross Entropy Loss Flat

C. *SHAP Library*

Deep learning model has been considered as not reliable as medical experts weigh in the fact that there is no explanation on the different features the model is using for its classification [23]. Recently, several libraries has been developed with the concept of explainable AI (XAI) such as Local Interpretable Model-agnostic Explanations (LIME) [24], Gradient-weighted class activation mapping (Grad-CAM)[25] and SHapley Additive exPlanations (SHAP) [26]. SHAP, compared to the other two, is not hard to interpret and is quite consistent in its ability to explain the pretrained model correctly [27]. The output of the SHAP image displays the actual image as well as highlighted areas in red and blue. Shades of red indicate elements that contributed favourably to the forecast of that specific class whereas shades of blue indicate parts that impacted adversely.

IV. RESULTS

We deployed the fine tuned XResNet50 model to predict unseen testing dataset to evaluate it, and the experimental findings are reported below. To begin, we examine the performance of our proposed model using a variety of metrics, including overall accuracy, precision, recall, and the F1-score. The ratio of successfully predicted observations to total observations for all kidney occurrences is denoted as the overall accuracy. Recall, often known as the sensitivity metric, represents the proportion of properly predicted instances in a given class to the total number of correctly predicted instances. Precision is described as the percentage of correctly predicted specific cases to all expected cases. The F1-score is the harmonic mean of recall and precision.

TABLE 2. MODEL PERFORMANCE METRICS

	Precision	Recall	F1-score
Normal	0.96	0.99	0.98
Stone	0.96	1.00	0.98
Tumor	0.98	0.90	0.93

During the training of the model, an early stopping occurred as the model reached its highest accuracy on epoch 9. The enhanced XResNet50 model achieved a training accuracy of 97% and a training loss of 0.11 while the model predicted correctly 97% of the testing data. Furthermore, from Table 2, the model was able to differentiate between a normal kidney and a kidney stone case (sensitivity of 0.99 for normal and 1.00 for Stone class) but it did not have the same effect in recognising tumor cases. It is further elaborated from Figure 3 that for the tumor class, 37 were classified as normal whereas 14 were predicted as Kidney stone and Figure 4 shows how high the probability the model predicted the wrong class for tumor cases. However, as all of the classifier assessment metrics are higher than or equal to 90 percent, we can clearly see the proposed model is a reliable model to differentiate a normal kidney to a kidney with a stone or tumor in it.

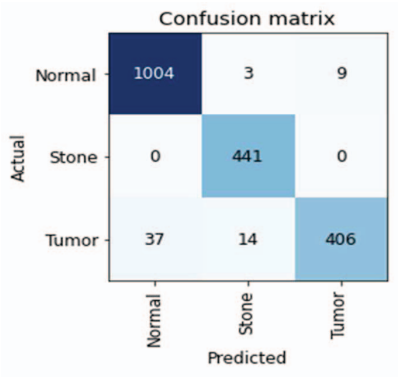

Fig. 3. Model confusion matrix

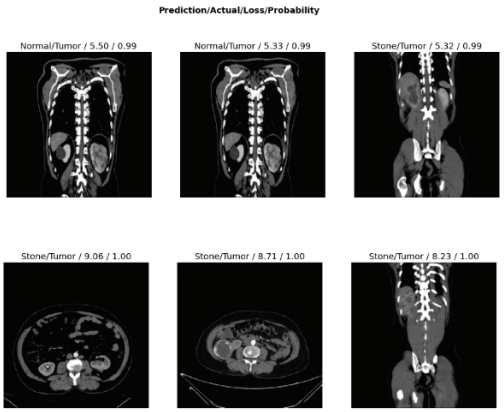

Fig. 4. Wrong Classification by model

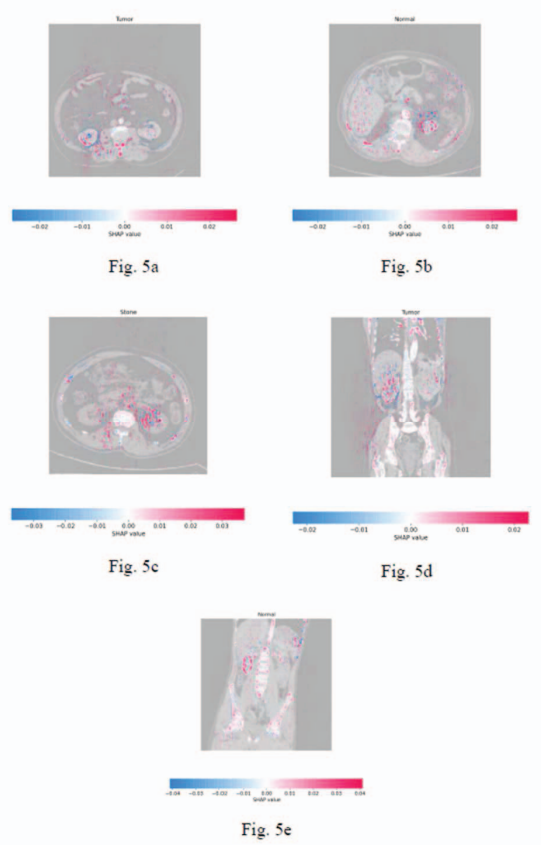

Figure 5. SHAP explanation

Although we obtained very good overall testing accuracy using our enhanced DNN model, we do not know how the model came to this result. Hence SHAP was selected for an in depth analysis of the model. SHAP, as described above, is a method that aids in the explanation of the output of our proposed XResNet50 model. Using SHAP, we can also investigate the important factors that the model has used to predict these 3 classes. From Figure 5, as an initial observation we can conclude that our model was mainly focusing on the kidney section of the CT-scan as most of the red and blue dots were concentrating around this specific region both in the axial and coronal planes. As in the case of Figure 5a and Figure 5d, since they were from the tumor class, we can note that despite that some of the red dots were in the kidney region, the model was not able to distinguish the tumor clearly as the focused regions were scattered. One of the main reasons is mainly due to some noise captured by the model which has affected its performance in distinguishing clearly the kidney tumor.

V. CONCLUSION

For this work, we have trained a fine-tuned XResNet50 model and classified normal (5,077 images), kidney stone (2167 images) and kidney tumor (2283 images) with a training accuracy and a testing accuracy of 97%. The model has correctly predicted all the images from the Stone class and has mispredicted 12 images of normal kidney and 51 images which has kidney tumor. From the SHAP images, it can be noted that for the Normal and Stone class, the model was focussing on the kidney section of the CT scan in both axial and coronal planes as most of the dots were present there. However for the Tumor class, despite that the model achieved a sensitivity of 90%, the regions where the model was focussing were scattered and not focussing on the kidney section of the CT scan. This is maybe due to some noise seen by the model which is affecting its performance.

REFERENCES

[1] National Kidney Foundation. Global Facts: About Kidney Disease | National Kidney Foundation [Internet]. 2015 [cited 2022 Jun 29]. Available from: https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease

[2] Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. Lancet. 2018 Nov 10;392(10159):2052–90.

[3] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394–424.

[4] Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2016: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2018 Nov 1;4(11):1553–68.

[5] Abufaraj M, Xu T, Cao C, Waldhoer T, Seitz C, D'andrea D, et al. Prevalence and Trends in Kidney Stone Among Adults in the USA: Analyses of National Health and Nutrition Examination Survey 2007-2018 Data. Eur Urol Focus. 2021 Nov;7(6):1468–75.

[6] Romero V, Akpinar H, Assimos DG. Kidney stones: a global picture of prevalence, incidence, and associated risk factors. Rev Urol. 2010;12(2–3):e86-96.

[7] Brikowski TH, Lotan Y, Pearle MS. Climate-related increase in the prevalence of urolithiasis in the United States. Proc Natl Acad Sci USA. 2008 Jul 15;105(28):9841–6.

[8] Kaufman J, Vicedo-Cabrera AM, Tam V, Song L, Coffel E, Tasian G. The impact of heat on kidney stone presentations in South Carolina under two climate change scenarios. Sci Rep. 2022 Jan 10;12(1):369.

[9] Yang C-W, Harris DCH, Luyckx VA, Nangaku M, Hou FF, Garcia Garcia G, et al. Global case studies for chronic kidney disease/end-stage kidney disease care. Kidney Int Suppl (2011). 2020 Mar;10(1):e24–48.

[10] Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak. 2021 Apr 10;21(1):125.

[11] Lee D, Yoon SN. Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges. Int J Environ Res Public Health. 2021 Jan 1;18(1).

[12] Ozrazgat-Baslanti T, Loftus TJ, Ren Y, Ruppert MM, Bihorac A. Advances in artificial intelligence and deep learning systems in ICU-related acute kidney injury. Curr Opin Crit Care. 2021 Dec 1;27(6):560–72.

[13] Alnazer I, Bourdon P, Urruty T, Falou O, Khalil M, Shahin A, et al. Recent advances in medical image processing for the evaluation of chronic kidney disease. Med Image Anal. 2021 Apr;69:101960.

[14] Yan X, Li X, Lu Y, Ma D, Mou S, Cheng Z, et al. Establishment and Evaluation of Artificial Intelligence-Based Prediction Models for Chronic Kidney Disease under the Background of Big Data. Evid Based Complement Alternat Med. 2022 Jul 8;2022:6561721.

[15] Sudharson S, Kokil P. An ensemble of deep neural networks for kidney ultrasound image classification. Comput Methods Programs Biomed. 2020 Dec;197:105709.

[16] Zhou L, Zhang Z, Chen Y-C, Zhao Z-Y, Yin X-D, Jiang H-B. A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors. Transl Oncol. 2019 Feb;12(2):292–300.

[17] Pedersen M, Andersen MB, Christiansen H, Azawi NH. Classification of renal tumour using convolutional neural networks to detect oncocytoma. Eur J Radiol. 2020 Dec;133:109343.

[18] Yildirim K, Bozdag PG, Talo M, Yildirim O, Karabatak M, Acharya UR. Deep learning model for automated kidney stone detection using coronal CT images. Comput Biol Med. 2021 Aug;135:104569.

[19] da Cruz LB, Araújo JDL, Ferreira JL, Diniz JOB, Silva AC, de Almeida JDS, et al. Kidney segmentation from computed tomography images using deep neural network. Comput Biol Med. 2020 Aug;123:103906.

[20] Sudharson S, Kokil P. Computer-aided diagnosis system for the classification of multi-class kidney abnormalities in the noisy ultrasound images. Comput Methods Programs Biomed. 2021 Jun;205:106071.

[21] Sudharson S, Kokil P. Abnormality Detection in the Renal Ultrasound Images using Ensemble MSVM Model. 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET). IEEE; 2019. p. 378–82.

[22] Islam MN, Hasan M, Hossain MK, Alam MGR, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto

detection of kidney cyst, stone and tumor from CT-radiography. Sci Rep. 2022 Jul 6;12(1):11440.

[23] Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging. 2020 Jun 20;6(6).

[24] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA: ACM Press; 2016. p. 1135–44.

[25] 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 2017 IEEE International Conference on Computer Vision (ICCV). IEEE; 2017. p. 618–26.

[26] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017.

[27] Sun J, Chakraborty T (Rohan), Noble J. A Comparative Study of Explainer Modules Applied to Automated Skin Lesion Classification. XI-ML@KI. 2020.