Development and multi-institutional validation of a deep learning model for grading of vesicoureteral reflux on voiding cystourethrogram: a retrospective multicenter study

Zhanchi Li,[a,k] Zelong Tan,[b,k] Zheyuan Wang,[c] Wenjuan Tang,[d] Xiang Ren,[d] Jinhua Fu,[d] Guangbing Wang,[e] Han Chu,[f] Jiarong Chen,[g] Yuhe Duan,[h] Likai Zhuang,[i,**] and Min Wu[j,*]

[a]Shanghai Jiao Tong University School of Medicine, Shanghai, China
[b]Department of Electronic Engineering, Tsinghua University, Beijing, China
[c]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[d]Department of Radiology, Shanghai Children's Hospital, School of Medicine, Shanghai Jiao Tong University, 355 Luding Road, Shanghai, 200062, China
[e]Department of Urology, Puyang People's Hospital, Henan, China
[f]Department of Urology, Anhui Provincial Children's Hospital, Anhui, China
[g]Department of Urology, The Children's Hospital of Guangxi Zhuang Autonomous Region, China
[h]Department of Urology, The Affiliated Hospital of Qingdao University, China
[i]Department of Urology, Children's Hospital of Fudan University, National Pediatric Medical Center of China, Shanghai, 201102, China
[j]Department of Urology, Shanghai Children's Hospital, School of Medicine, Shanghai Jiao Tong University, 355 Luding Road, Shanghai, 200062, China

eClinicalMedicine
2024;69: 102466
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102466

Summary

Background Voiding cystourethrography (VCUG) is the gold standard for the diagnosis and grading of vesicoureteral reflux (VUR). However, VUR grading from voiding cystourethrograms is highly subjective with low reliability. This study aimed to develop a deep learning model to improve reliability for VUR grading on VCUG and compare its performance to that of clinicians.

Methods In this retrospective study in China, VCUG images were collected between January 2019 and September 2022 from our institution as an internal dataset for training and 4 external data sets as external testing set for validation. Samples were divided into training (N = 1000) and validation sets (N = 500), internal testing set (N = 168), and external testing set (N = 280). An ensemble learning-based model, Deep-VCUG, using Res-Net 101 and the voting methods was developed to predict VUR grade. The grading performance was assessed using heatmaps, area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, and F1 score in the internal and external testing set. The performances of four clinicians (2 pediatric urologists and 2 radiologists) with and without the Deep-VCUG assisted to predict VUR grade were explored in external testing sets.

Findings A total of 1948 VCUG images were collected (Internal dataset = 1668; multi-center external dataset = 280). For assessing unilateral VUR grading, the Deep-VCUG achieved AUCs of 0.962 (95% confidence interval [CI]: 0.943–0.978) and 0.944 (95% [CI]: 0.921–0.964) in the internal and external testing sets, respectively, for bilateral VUR grading, the Deep-VCUG also achieved high AUCs of 0.960 (95% [CI]: 0.922–0.983) and 0.924 (95% [CI]: 0.887–0.957). The Deep-VCUG model using voting method outperformed single model and clinician in terms of classification based on VCUG image. Moreover, Under the Dee-VCUG assisted, the classification ability of junior and senior clinicians was significantly improved.

Interpretation The Deep-VCUG model is a generalizable, objective, and accurate tool for vesicoureteral reflux grading based on VCUG imaging and had good assistance with clinicians to VUR grading applicability.

Funding This study was supported by Natural Science Foundation of China, "Fuqing Scholar" Student Scientific Research Program of Shanghai Medical College, Fudan University, and the Program of Greater Bay Area Institute of Precision Medicine (Guangzhou).

*Corresponding author.
**Corresponding author.
E-mail addresses: doctorwumin@163.com (M. Wu), lszx04336@163.com (L. Zhuang).
[k]These authors have contributed equally to this work.

Keywords: Deep learning; Voiding cystourethrography; Bilateral vesicoureteral reflux; Ensemble learning

Research in context

Evidence before this study

We searched PubMed and Web of Science with the terms "(Vesicoureteral Reflux OR VUR)" AND "(Voiding Cystourethrogram OR VCUG)" AND "(deep learning OR artificial intelligence)" for papers published from database inception to Nov 7, 2023, with no language restrictions. We find that 2 studies are based on machine learning. Only 1 research about the deep learning-based classification of VUR using VCUG was published. However, these studies have various limitations, including small sample size, single-center design, and still relied on time-consuming manual delineation. In addition, these methods are not for bilateral VUR grading.

Added value of this study

Our study proposed a deep learning model, Deep-VCUG, on a large multi-institutional cohort, for VUR grade based on a single VUCG image. We evaluated it in an internal test set and multi-center set and compared it with junior and senior clinicians. Moreover, Deep-VCUG-assisted strategies could improve clinicians' s performance.

Implications of all the available evidence

Our findings show that the Deep-VCUG has an excellent ability to predict VUR grade based on VCUG images, which performs equivalent to or even better than senior experts and can assist in improving the performance of Junior clinicians. In the future, more prospective multicenter validation will provide strong evidence for the performance of our Deep-VCUG in assisting the clinician.

Introduction

Vesicoureteral reflux (VUR) is the most common urinary tract disease in children with recurrent urinary tract infections in childhood. Voiding cystourethrography (VCUG) is the gold standard for the diagnosis and grading of VUR.[1,2] However, VUR grading from voiding cystourethrograms is highly subjective with low reliability.[3,4] Therefore, there is a need to develop a standardized, objective, and reliable method to accurately grade VUR severity that improves reliability for VUR grading.

Recently, with the gradual development of Artificial intelligence (AI) in biomedical applications. Machine learning (ML) and deep learning (DL) has been proposed as a potential solution for challenges with VUR grading.[5–7] But ML methods have needed to manually mark these features: the ureteropelvic junction (UPJ), the ureterovesical junction (UVJ), ureter width, and tortuosity[5,6]; the DL model is trained and evaluated on single-center datasets.[7] Meanwhile, a new guideline has been designed explicitly for urology. The Standardized Reporting of Machine Learning Applications in Urology (STREAM-URO) framework provides a list of minimum reporting standards for artificial intelligence studies in urology that promotes more transparent and high-quality ML studies in the urological community.[5,8]

Herein, the objective of this study is to develop and validate a deep learning network model (Deep-VCUG) with ensemble learning for automatic VUR grading from VCUG images in accordance with STREAM-URO. The primary outcome of this study was the assessment of the performance of Deep-VCUG model. Secondary outcomes were comparing its performance to that of clinicians.

Methods

Ethics statement

This study was approved by the research ethics committee of the Fudan Children's Hospital (Ref. No.201756). An informed consent exemption from the Institutional Review Board was obtained. This study was conducted in accordance with the STREAM-URO framework (details in Supplementary Appendix Table S1). Computer codes are available online (https://github.com/Li-zhanchi/VCUG.git).

Datasets

This retrospective study analyzed 2440 VCUG images from the imaging archive, a retrospective, observational study. Participants were recruited from January 2019 to September 2022 and formed a consecutive series. The raw VCUG images were collected for all patients with reflux, regardless of indication, along with age at imaging, sex, indication for VCUG, and radiologist-reported VUR grade.

Each abstracted image was taken at the peak of the filling phase to reduce variability and ensure adequate contrast for annotation.[5] A total of 772 participants were excluded because of poor image quality, excessive malformation (e.g., cloacal m malformation, ectopic ureter, and hypospadias), lack of anteroposterior view, or other reasons (e.g., medical equipment obstructing the view,

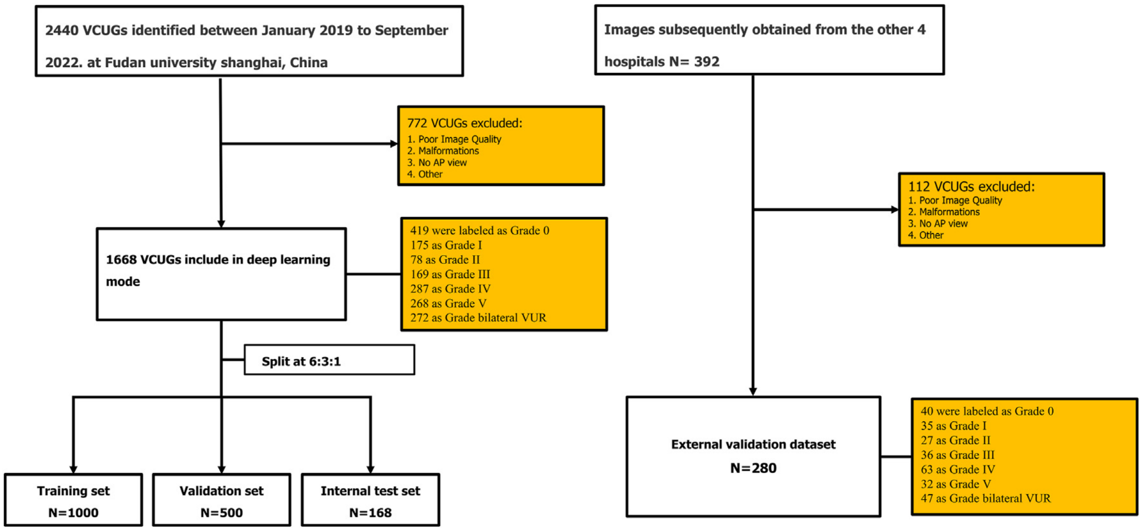

Fig. 1: Flowchart showing participant selection from the VCUG database.

post-dextranomer hyaluronate acid, post-surgical correction) (Fig. 1).

We used four independent hospitals' VCUG datasets as external testing cohorts to evaluate Deep-VCUG: Puyang People's Hospital, (Henan, China); Anhui Provincial Children's Hospital (Anhui, China); The Children's Hospital of Guangxi Zhuang autonomous region, (Guangxi, China); The Affiliated Hospital of Qingdao University, (Shandong, China) and data are described in Supplementary Appendix Table S2.

Image readings

The ground truth of VUR grade was analyzed using the International Reflux Society classification criteria.[9] Three clinicians' experts were invited to identify the ground truth of the internal and external datasets, which included two pediatric urologists (M.W. and L.K.Z. each with 10 years of experience reading VCUG radiographs, respectively) and an associate chief physician of pediatric radiology (W.J.T. with 20 years of experience in reading VCUG radiographs.). Readers were blinded to clinical information and other imaging results. Initial training was performed by pediatric radiology (W.J.T.). For the VUR grading readings, two pediatric urologists (M.W. and L.K.Z) read all VCUG radiographs (including internal and external datasets) independently. In cases of agreement, these readings served as ground truth; in cases of disagreement, readings were adjudicated by a radiologist (W.J.T.) to establish ground truth.

Data processing

All VCUG images with the entire bladder and ureter area were cropped and then scaled to size 512 × 512. Then, all individual scans had been normalized to the 0~1 interval by subtracting the mean value and dividing by the variance. Finally, the training set was augmented by random affine transformations with rotation, translation, scaling, and random perspective transformations.

Model architecture

In this study, a deep learning model called Deep-VCUG was developed by combining deep convolutional neural networks (CNNs) and ensemble learning. The model architecture consists of two main parts (Fig. 2): The first component is a VUR lateral classification model, which uses the ResNet-101 architecture to identify the input VCUG image as a unilateral or bilateral VUR. The second component predicts VUR grade using an ensemble learning method based on the unilateral or bilateral results of VUR from the first component. The ensemble learning approach involves training five weak models and combining their predictions using the voting method to obtain a final result (Detail described in Supplementary Method).

Model training

During model training, the networks were trained for 300 epochs by the Adam optimizer with a learning rate of 1×10^{-3}, decaying by 0.1 after 90 for each epoch. All models were trained using a GTX TITAN X with 12G and used the Pytorch (Pytorch version 1.8, Python version 3.7).

Model performance and compares with other models

In the internal and external testing sets, the performance of Deep-VCUG model was compared with that of the other models, which included MobilNetv2, GoogLeNet, ResNet 101, DenseNet161, and EfficientNet-B0. The receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, and specificity are used to compare.

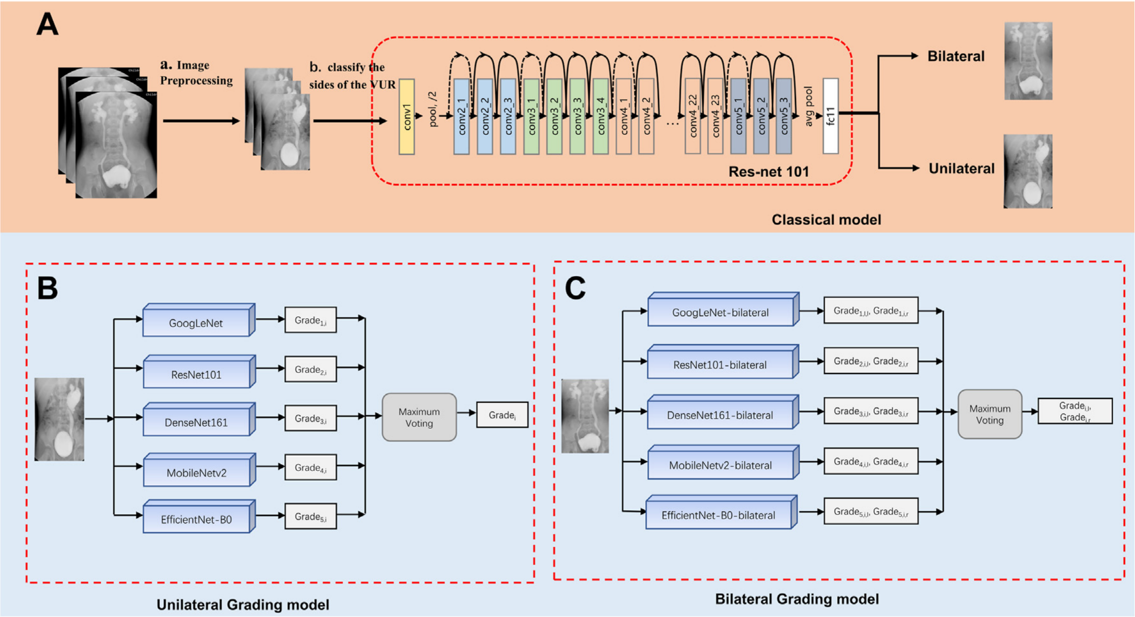

Fig. 2: The multistep cascade experiment pathway of Deep-VCUG. (A) The process of classifying the lateral of the VUR. First, Image Preprocessing (a). Second, classify the sides of the VUR using ResNet-101 (b). (B) The process of unilateral grading classifying images by ensemble learning model (C) The process of bilateral grading classifying images by ensemble learning model. After obtaining the unilateral or bilateral results of VUR, five classification models were trained and combined to obtain an advanced classification model. The final classification result through the voting method.

Clinicians' performance with and without Deep-VCUG assistance

To evaluate the efficacy of the Deep-VCUG model in the clinical. VCUG images from external testing sets were used to assess the radiologist's performance of with and without the Deep-VCUG assistance. Two radiologists [with professional experience of 5 years (X.R.) and 10 years (C.H.F.), respectively] and two pediatric urologists [(with professional experience of 6 years (G.B.W.) and 10 years (R.J.C), respectively)]] are invited. Only VCUG images were available to four clinicians. First, four clinicians assessed the grading of VUR on VCUG image according to International Reflux Society classification criteria.[9] Then, they are performed to reassess VUR grade based on the predicted probability of lesions provided by Deep-VCUG. They could choose either not to change or else to read just their first result. In addition, the agreement between all pairs of clinicians with and without Deep-VCUG assistance was calculated using Cohen's kappa value.

Visualization and application of DL model for representative cases

For investigating the interpretability of the Deep-VCUG, the network was visualized by applying the Gradient-weighted Class Activation Mapping (Grad-CAM),[10] which could produce a coarse localization map highlighting the import regions for classification target. In addition, we have selected some representative cases to demonstrate the visualization capabilities of the DL model.

Statistical analysis

Continuous variables are recorded as appropriate mean ± standard deviation (std) and categorical variables are recorded as numbers and percentages. The performance of all models used for the classification tasks was assessed in terms of the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, F1-value, Precision, and accuracy. The *DeLong test* was used to compare differences between AUCs. In contrast, the *Wilcoxon test* was to compare the differences in sensitivity, specificity, F1 score, precious, and accuracy. Statistical analyses in this study were conducted using the Python 3.8 programming language (https://www.python.org/) with key libraries including "SciPy" (https://www.scipy.org/), "scikit-learn" (https://scikit-learn.org/), "matplotlib" (https://matplotlib.org/), and other pertinent packages. A bilateral *P* value of less than 0.05 was deemed to indicate statistical significance.

Role of the funding source

The source of funding played no role in the design of the study protocol, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the manuscript for publication. The corresponding author had full access to all data in the study and assumes final responsibility for the decision to submit the manuscript for publication. All authors approved the decision to submit.

Results

Clinical characteristics

A total of 2440 VCUG images were identified in the study period. Of these, 1668 VCUGs meeting eligibility criteria (mean age, 41.57 ± 42.23 months; 927 males (55.58%) and 741 females (44.42%)); The participants were split by using split-sample validation into training, validation, and internal testing sets of 60% (n = 1000), 30% (n = 500), and 10% (n = 168), respectively. In the external testing set, participants were mean age of 23.13 ± 28.20 months; participants were 165 males (58.93%) and 115 females (41.07%). Table 1 provides an overview of participant characteristics.

Performance of the Deep-VCUG and compared with subnetwork model

As shown in Fig. 3 and Table 2. In the internal test set, for unilateral and bilateral VUR grade categorization, the Deep-VCUG achieved AUCs of 0.940 (95% CI: 0.894–0.987) and 0.891 (95% CI: 0.816–0.945), sensitivity of 0.940 (95% CI: 0.894–0.987) and 0.891 (95% CI: 0.816–0.945), and specificity of 0.940 (95% CI: 0.894–0.987) and 0.891 (95% CI: 0.816–0.945) respectively. The AUC of Deep-VCUG was statistically higher than other model ($P < 0.05$), except for Mobildenet_v2 of unilateral VUR ($P = 0.0786$) and resnet 101 ($P = 0.892$) and densenet161 ($P = 0.1904$) of bilateral VUR. In the external testing set, for unilateral and bilateral VUR grade categorization, the Deep-VCUG achieved AUCs of 0.944 (95% CI: 0.921–0.964) and 0.924 (95% CI: 0.887–0.957), sensitivity of 0.807 (95% CI: 0.755–0.853) and 0.745 (95% CI: 0.660–0.830), and specificity of 0.958 (95% CI: 0.946–0.970) and 0.808 (95% CI: 0.738–0.874), The AUCs of Deep-VCUG was higher than other models, except for MobileNetv2, GoogLeNet for unilateral VUR and MobileNetv2 for bilateral VUR. The AUCs of Deep-VCUG are not statistically different ($P > 0.05$), except for Efficientnet_B0 of unilateral VUR and GoogleNet, Resnet 101, and Efficientnet_B0 of bilateral VUR ($P < 0.05$). The detailed performance of the Deep-VCUG and other models is summarized in Supplementary Appendix Tables S2–S6.

Performance of the Deep-VCUG and comparison with clinicians

The sensitivity and specificity points of the four clinicians in the external testing sets are drawn on the same ROC curve in (Fig. 4). (Fig. 5) are the confusion matrices that Deep-VCUG mode and four clinicians in the external testing data. For unilateral VUR (Table 3), the Deep-VCUG model demonstrated an accuracy of 0.807, specificity of 0.958, precision of 0.827, sensitivity of 0.807, and F1 score of 0.807. The AUC was 0.944 (95% CI: 0.921–0.964) which was the highest in comparison to that of the four clinicians (all $P < 0.05$). The AUCs of senior pediatric urologist and senior radiologist [0.809 (95% CI: 0.750–0.821); 0.822 (95% CI: 0.766–0.840), respectively] were significantly higher ($P < 0.05$) than that of junior pediatric urologist and junior radiologist [0.722 (95% CI: 0.729–0.804); 0.798 (95% CI: 0.745–0.821), respectively]. Additionally, the sensitivity and specificity of the DL model (0.807, 0.958) were higher than four clinicians, but were non-significantly different from that of two senior clinicians (0.682 and 0.703 of sensitivity, 0.922 and 0.924 of specificity, respectively) and two junior clinicians (0.621 and 0.624 of sensitivity, 0.908 and 0.920 of specificity, respectively) ($P > 0.05$).

For bilateral VUR (Table 3), the Deep-VCUG model achieved an accuracy of 0.745, a precious of 0.766, a specificity of 0.808, F1 score of 0.720. The AUC was

Participant Characteristics	Training Set (N = 1000)	Validation Set (N = 500)	Internal test sets (N = 168)	External Test Set (N = 280)
Age (month, mean ± SD)[a]	39.77 ± 41.74 (1–179)	41.98 ± 40.14 (1–177)	51.1 ± 49.53 (1–180)	23.13 ± 28.20 (1–177)
Sex				
Male	538 (53.80%)	317 (63.40%)	72 (42.86%)	165 (58.93%)
Female	462 (46.20%)	183 (36.60%)	96 (57.14%)	115 (41.07%)
VUR Grading				
Grade-0	251 (25.10%)	126 (25.20%)	42 (25.00%)	40 (14.29%)
Grade-1	105 (10.50%)	52 (10.40%)	18 (10.71%)	35 (12.50%)
Grade-2	47 (4.70%)	23 (4.60%)	9 (5.36%)	27 (9.64%)
Grade-3	101 (10.10%)	51 (10.20%)	17 (10.21%)	36 (12.86%)
Grade-4	172 (17.20%)	86 (17.20%)	29 (17.26%)	63 (22.50%)
Grade-5	161 (16.10%)	80 (16.00%)	27 (16.07%)	32 (11.43%)
Bilateral VUR	163 (16.30%)	82 (16.40%)	27 (16.07%)	47 (16.79%)

N = Number of VUCG images. Data are presented as n (%); SD = standard deviation. [a]Data are mean ± SD, with the range in parentheses.

Table 1: **Demographic Characteristics from the Data Sets from the VUCG Initiative and the external test set.**

	Accuracy (95% CI)	Precision (95% CI)	Sensitivity (95% CI)	F1 Score (95% CI)	AUROC (95% CI)	Specificity (95% CI)
Unilateral VUR grading (N = 374)						
Internal testing set (N = 141)						
Deep-VCUG	0.805 (0.738–0.866)	0.761 (0.688–0.841)	0.805 (0.738–0.866)	0.782 (0.710–0.851)	0.962 (0.943–0.978)	0.960 (0.941–0.975)
MobileNetv2	0.779 (0.711–0.846)	0.749 (0.683–0.828)	0.779 (0.711–0.846)	0.762 (0.695–0.832)	0.948 (0.920–0.972)	0.957 (0.941–0.972)
GoogLeNet	0.597 (0.523–0.671)	0.518 (0.417–0.633)	0.597 (0.523–0.671)	0.529 (0.438–0.618)	0.881 (0.845–0.914)	0.878 (0.856–0.900)
ResNet 101	0.738 (0.671–0.805)	0.729 (0.652–0.809)	0.738 (0.671–0.805)	0.731 (0.658–0.800)	0.931 (0.899–0.962)	0.951 (0.935–0.968)
DenseNet161	0.758 (0.691–0.826)	0.783 (0.724–0.850)	0.758 (0.691–0.826)	0.762 (0.693–0.826)	0.942 (0.916–0.967)	0.955 (0.938–0.970)
EfficientNet-B0	0.570 (0.483–0.644)	0.539 (0.438–0.656)	0.570 (0.483–0.644)	0.533 (0.444–0.616)	0.900 (0.868–0.928)	0.893 (0.871–0.915)
External testing set (N = 233)						
Deep-VCUG	0.807 (0.755–0.858)	0.827 (0.788–0.873)	0.807 (0.755–0.858)	0.807 (0.756–0.858)	0.944 (0.921–0.964)	0.958 (0.946–0.970)
MobileNetv2	0.785 (0.734–0.837)	0.792 (0.747–0.845)	0.785 (0.734–0.837)	0.786 (0.734–0.838)	0.952 (0.933–0.968)	0.954 (0.942–0.965)
GoogLeNet	0.790 (0.734–0.841)	0.804 (0.760–0.852)	0.790 (0.734–0.841)	0.790 (0.736–0.839)	0.954 (0.935–0.971)	0.952 (0.938–0.964)
ResNet 101	0.764 (0.704–0.820)	0.782 (0.731–0.834)	0.764 (0.704–0.820)	0.765 (0.706–0.820)	0.933 (0.907–0.953)	0.950 (0.936–0.962)
DenseNet161	0.768 (0.717–0.820)	0.792 (0.751–0.839)	0.768 (0.717–0.820)	0.772 (0.718–0.823)	0.943 (0.924–0.963)	0.948 (0.934–0.961)
EfficientNet-B0	0.326 (0.266–0.391)	0.295 (0.234–0.368)	0.326 (0.266–0.391)	0.290 (0.231–0.356)	0.725 (0.684–0.768)	0.837 (0.806–0.863)
Bilateral VUR grading (N = 74)						
Internal testing set (N = 27)						
Deep-VCUG	0.796 (0.685–0.889)	0.833 (0.671–0.917)	0.796 (0.685–0.889)	0.775 (0.640–0.879)	0.960 (0.922–0.983)	0.936 (0.898–0.969)
MobileNetv2	0.741 (0.630–0.852)	0.782 (0.604–0.882)	0.741 (0.630–0.852)	0.720 (0.585–0.847)	0.936 (0.888–0.975)	0.911 (0.861–0.954)
GoogLeNet	0.741 (0.630–0.852)	0.766 (0.590–0.869)	0.741 (0.630–0.852)	0.716 (0.581–0.840)	0.932 (0.876–0.971)	0.912 (0.863–0.951)
ResNet 101	0.741 (0.611–0.852)	0.804 (0.647–0.886)	0.741 (0.611–0.852)	0.721 (0.583–0.847)	0.960 (0.925–0.984)	0.904 (0.851–0.950)
DenseNet161	0.778 (0.667–0.889)	0.795 (0.657–0.913)	0.778 (0.667–0.889)	0.762 (0.632–0.880)	0.943 (0.899–0.980)	0.946 (0.910–0.976)
EfficientNet-B0	0.667 (0.537–0.778)	0.679 (0.560–0.817)	0.667 (0.537–0.778)	0.656 (0.522–0.777)	0.919 (0.863–0.959)	0.893 (0.836–0.935)
External testing set (N = 47)						
Deep-VCUG	0.745 (0.660–0.830)	0.766 (0.671–0.852)	0.745 (0.660–0.830)	0.720 (0.617–0.819)	0.924 (0.887–0.957)	0.808 (0.738–0.874)
MobileNetv2	0.723 (0.628–0.809)	0.740 (0.640–0.827)	0.723 (0.628–0.809)	0.699 (0.587–0.788)	0.930 (0.888–0.963)	0.799 (0.728–0.859)
GoogLeNet	0.628 (0.521–0.734)	0.646 (0.543–0.760)	0.628 (0.521–0.734)	0.616 (0.510–0.724)	0.878 (0.834–0.918)	0.787 (0.713–0.850)
ResNet 101	0.691 (0.606–0.777)	0.691 (0.592–0.791)	0.691 (0.606–0.777)	0.676 (0.577–0.773)	0.900 (0.858–0.937)	0.803 (0.728–0.864)
DenseNet161	0.670 (0.574–0.766)	0.688 (0.582–0.787)	0.670 (0.574–0.766)	0.639 (0.527–0.742)	0.916 (0.881–0.949)	0.769 (0.697–0.835)
EfficientNet-B0	0.489 (0.383–0.596)	0.511 (0.407–0.634)	0.489 (0.383–0.596)	0.493 (0.391–0.604)	0.843 (0.795–0.889)	0.730 (0.644–0.805)

Table 2: The performance of Deep-VCUG in the internal and external testing sets for unilateral and bilateral VUR grading.

0.924 (95% CI: 0.887–0.957) which was the highest in comparison to that of the four clinicians (all $P < 0.05$). The AUC of senior pediatric urologist [0.796 (95% CI: 0.714–0.835)] was significantly higher ($P < 0.05$) than that of junior pediatric urologist [0.732 (95% CI: 0.650–0.773)], senior radiologist [0.821 (95% CI: (0.740–0.863)] was not significantly different ($P = 0.0578$) than that of junior radiologist [0.783 (95% CI: 0.701–0.821)]. Moreover, the sensitivity and specificity of the DL model (0.745, 0.808) were higher than four clinicians, but were non-significantly different from that of two senior clinicians (0.660 and 0.702 of sensitivity, 0.859 and 0.862 of specificity, respectively) and two junior clinicians (0.553 and 0.638 of sensitivity, 0.806 and 0.837 of specificity, respectively)) ($P > 0.05$). Appendix Tables S4–S7 show the contrast P-value of AUC, accuracy, sensitivity, and specificity.

Clinician's performance with Deep-VCUG assistance

The detailed changes in each diagnostic index of four clinicians with and without the aid of the Deep-VCUG model are shown in Table 3. In the external test set, except for the senior pediatric urologist who graded bilateral VUR grading ($P = 0.1815$), the AUCs of remaining clinicians for VUR grading were significantly higher than the corresponding previous values ($P < 0.05$). Furthermore, for unilateral and bilateral VUR grading, the junior clinician's average AUC improved by 0.053 and 0.054, and the senior clinician's AUC improved by 0.051 and 0.029 on average. Moreover, the average agreement among all clinicians with the assistance of the Deep-VCUG for unilateral and bilateral VUR grading increased from 0.06 to 0.17 and from 0.01 to 0.13, respectively. Detailed results can be found in Fig. 6.

Model interpretation and visualization

To better understand whether the CNN focuses on the suitable area, we employed class activation mapping (CAM) to visualize the internal features of the neural network. We found that the neural network focused on the region of the ureter in grading VUR. These findings indicate that the model learned to assess the correct

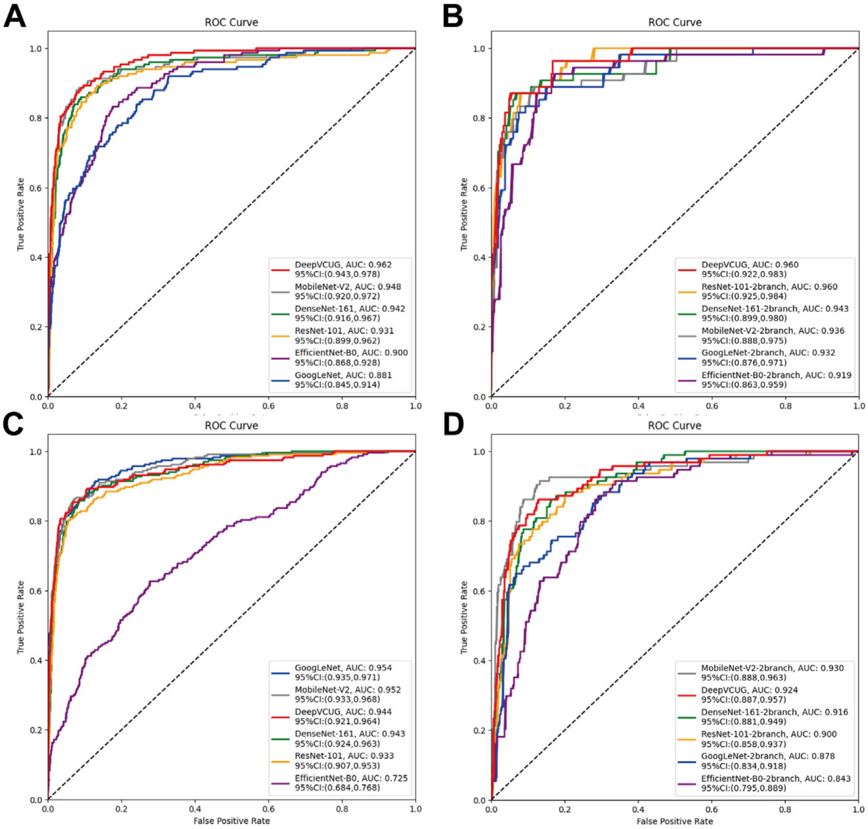

Fig. 3: Performance of the Deep-VCUG model and other models in the internal and external testing set. (A) Unilateral VUR and (B) bilateral VUR performance on the internal test set. (C) Unilateral VUR and (D) bilateral VUR performance on external data set.

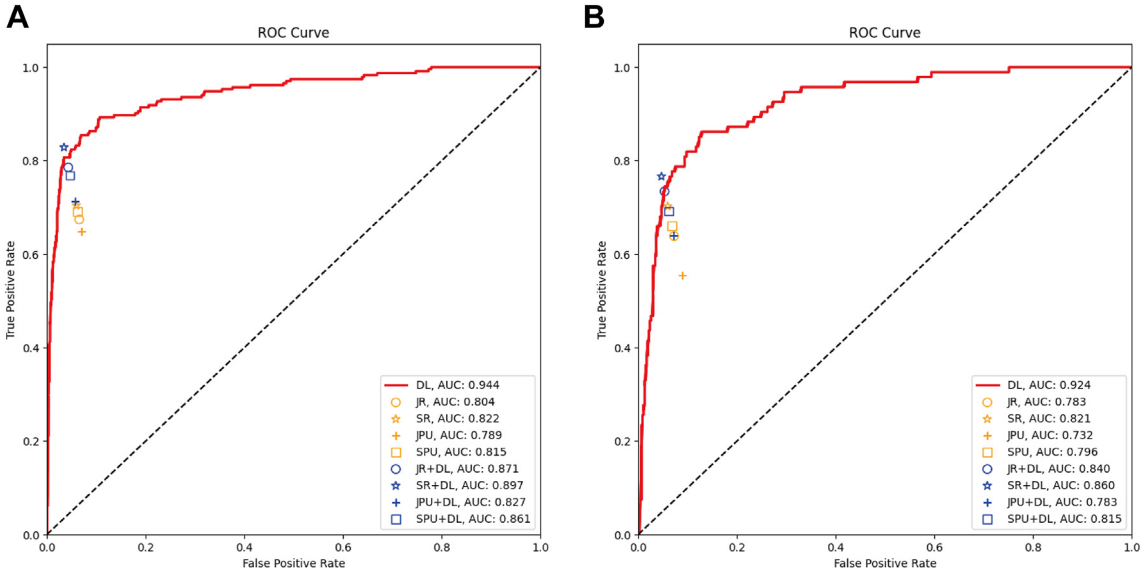

Fig. 4: The performance of the Deep-VCUG and clinicians with and without the Deep-VCUG assist on the external testing set for unilateral (A) and bilateral VUR grading (B). ROC, receiver operating characteristic; AUC, area under receiver operating characteristic curve; DL, Deep-VCUG model; JPU, Junior Pediatric urologist; SPU, Senior Pediatric urologist; JR, Junior radiologist; SR, Senior radiologist.

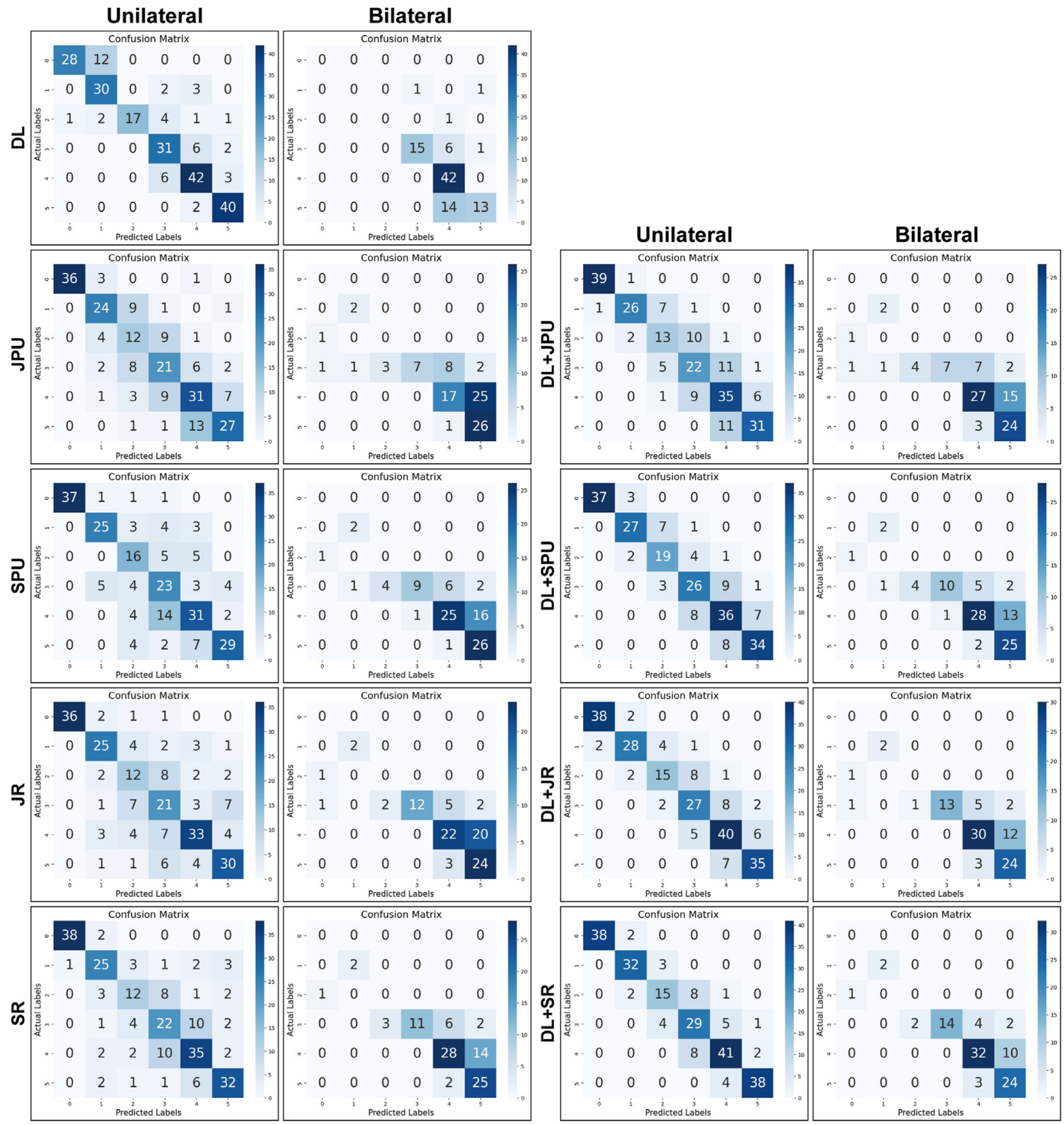

Fig. 5: Confusion matrices of the Deep-VCUG model and four clinicians with and without Deep-VCUG assist in the external testing set. The color depends on the number inside the square: the higher the number, the darker the color. DL, Deep-VCUG model; JPU, Junior Pediatric urologist; SPU, Senior Pediatric urologist; JR, Junior radiologist; SR, Senior radiologist.

features instead of learning image correlations. Patient examples for the actual use of the established CNN model are displayed in Fig. 7.

Discussion

In this study, we developed a Deep-VCUG model to classify the grading of unilateral VUR from VCUG images. For unilateral VUR from VCUG images, the model achieved AUCs of 0.962 and 0.944 in internal and external testing sets, respectively. Similarly, the model also performed well in classifying the grading of bilateral VUR from VCUG images, with AUCs of 0.960 and 0.924 in internal and external testing sets. Moreover, Deep-VCUG assisted can improve clinicians' performance. As far as we know, our study developed and validated a deep learning approach for VUR grading on a large multi-institutional cohort with high accuracy and reliability efficiency, which can accurately predict VUR grade and assist clinicians to improving the reliability.

	Accuracy (95% CI)	Precision (95% CI)	Sensitivity (95% CI)	F1 Score (95% CI)	AUROC (95% CI)	Specificity (95% CI)
Unilateral VUR (N = 233)						
DL	0.807 (0.755–0.858)	0.827 (0.788–0.873)	0.807 (0.755–0.858)	0.807 (0.756–0.858)	0.944 (0.921–0.964)	0.958 (0.946–0.970)
JPU	0.621 (0.588–0.708)	0.631 (0.610–0.732)	0.621 (0.588–0.708)	0.620 (0.595–0.715)	0.772[a] (0.729–0.804)	0.908 (0.913–0.944)
SPU	0.682 (0.635–0.747)	0.690 (0.662–0.770)	0.682 (0.635–0.747)	0.683 (0.644–0.753)	0.809[a] (0.758–0.829)	0.922 (0.922–0.951)
JR	0.664 (0.614–0.734)	0.676 (0.633–0.754)	0.664 (0.614–0.734)	0.665 (0.618–0.740)	0.798[a] (0.745–0.821)	0.920 (0.920–0.951)
SR	0.703 (0.648–0.764)	0.700 (0.655–0.768)	0.703 (0.648–0.764)	0.700 (0.649–0.762)	0.822[a] (0.766–0.840)	0.924 (0.922–0.952)
JPU + DL	0.691 (0.657–0.773)	0.691 (0.673–0.784)	0.691 (0.657–0.773)	0.689 (0.659–0.773)	0.815[a] (0.771–0.846)	0.919 (0.921–0.954)
SPU + DL	0.746 (0.717–0.824)	0.751 (0.728–0.833)	0.746 (0.717–0.824)	0.745 (0.720–0.825)	0.848[a] (0.810–0.880)	0.934 (0.936–0.964)
JR + DL	0.771 (0.734–0.837)	0.770 (0.738–0.842)	0.771 (0.734–0.837)	0.768 (0.731–0.837)	0.862[a] (0.821–0.888)	0.938 (0.939–0.965)
SR + DL	0.810 (0.777–0.880)	0.809 (0.785–0.884)	0.810 (0.777–0.880)	0.809 (0.778–0.880)	0.886[a] (0.849–0.917)	0.951 (0.952–0.976)
Bilateral VUR (N = 47)						
DL	0.745 (0.660–0.830)	0.766 (0.671–0.852)	0.745 (0.660–0.830)	0.720 (0.617–0.819)	0.924 (0.887–0.957)	0.808 (0.738–0.874)
JPU	0.553 (0.457–0.660)	0.681 (0.610–0.770)	0.553 (0.457–0.660)	0.540 (0.435–0.659)	0.732[a] (0.650–0.773)	0.806 (0.749–0.861)
SPU	0.660 (0.564–0.755)	0.744 (0.660–0.829)	0.660 (0.564–0.755)	0.661 (0.559–0.760)	0.796[a] (0.714–0.835)	0.859 (0.802–0.902)
JR	0.638 (0.543–0.734)	0.733 (0.660–0.820)	0.638 (0.543–0.734)	0.648 (0.554–0.748)	0.783[a] (0.701–0.821)	0.837 (0.772–0.890)
SR	0.702 (0.606–0.798)	0.778 (0.716–0.854)	0.702 (0.606–0.798)	0.709 (0.615–0.802)	0.821[a] (0.740–0.863)	0.862 (0.803–0.913)
JPU + DL	0.638 (0.543–0.734)	0.742 (0.677–0.823)	0.638 (0.543–0.734)	0.638 (0.538–0.745)	0.783[a] (0.701–0.821)	0.841 (0.779–0.895)
SPU + DL	0.691 (0.606–0.787)	0.764 (0.690–0.853)	0.691 (0.606–0.787)	0.698 (0.611–0.794)	0.815[a] (0.740–0.856)	0.872 (0.819–0.922)
JR + DL	0.734 (0.638–0.819)	0.789 (0.725–0.863)	0.734 (0.638–0.819)	0.742 (0.653–0.828)	0.840[a] (0.760–0.877)	0.871 (0.816–0.922)
SR + DL	0.766 (0.670–0.840)	0.813 (0.745–0.884)	0.766 (0.670–0.840)	0.775 (0.686–0.854)	0.860[a] (0.780–0.891)	0.888 (0.830–0.938)

Abbreviations: JPU, Junior Pediatric urologist; SPU, Senior Pediatric urologist; JR, Junior radiologist; SR, Senior radiologist; DL, Deep-VCUG model. [a]The differences between clinicians and DL model, and the differences among clinicians were compared, *P* values were calculated. Detailed results are presented in Appendix Tables S7–S18; *P* < 0.05, Significant difference with the DL model.

Table 3: **The performance of Deep-VCUG, clinicians alone, and Deep-VCUG-assisted clinicians in external dataset.**

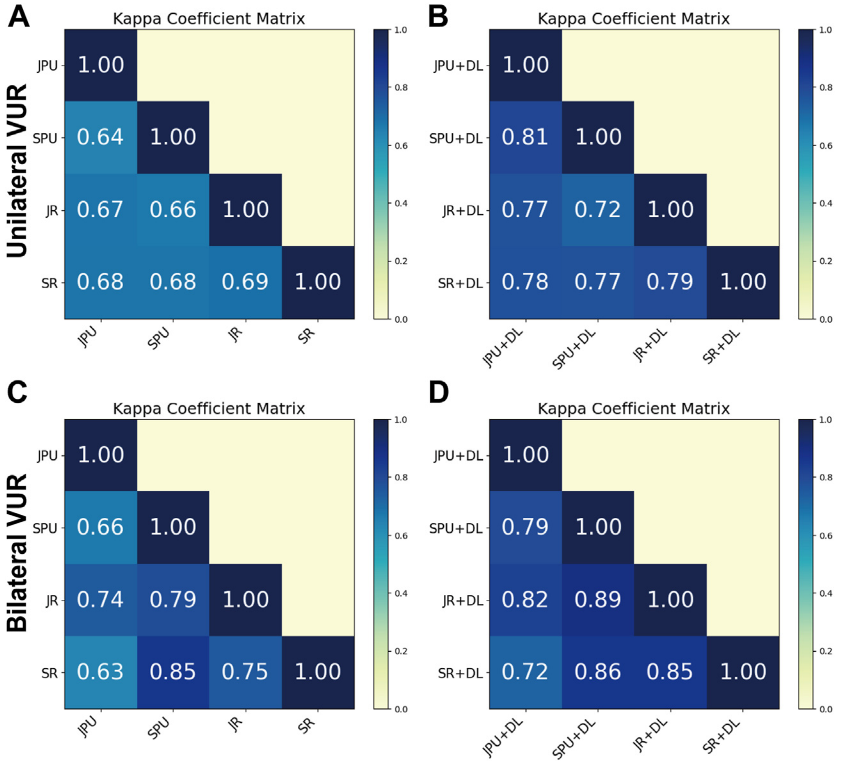

Fig. 6: The agreement degree of pairs of clinicians without and with Deep-VCUG assistance of unilateral VUR (A, B) and bilateral VUR (C, D) in the external testing set.

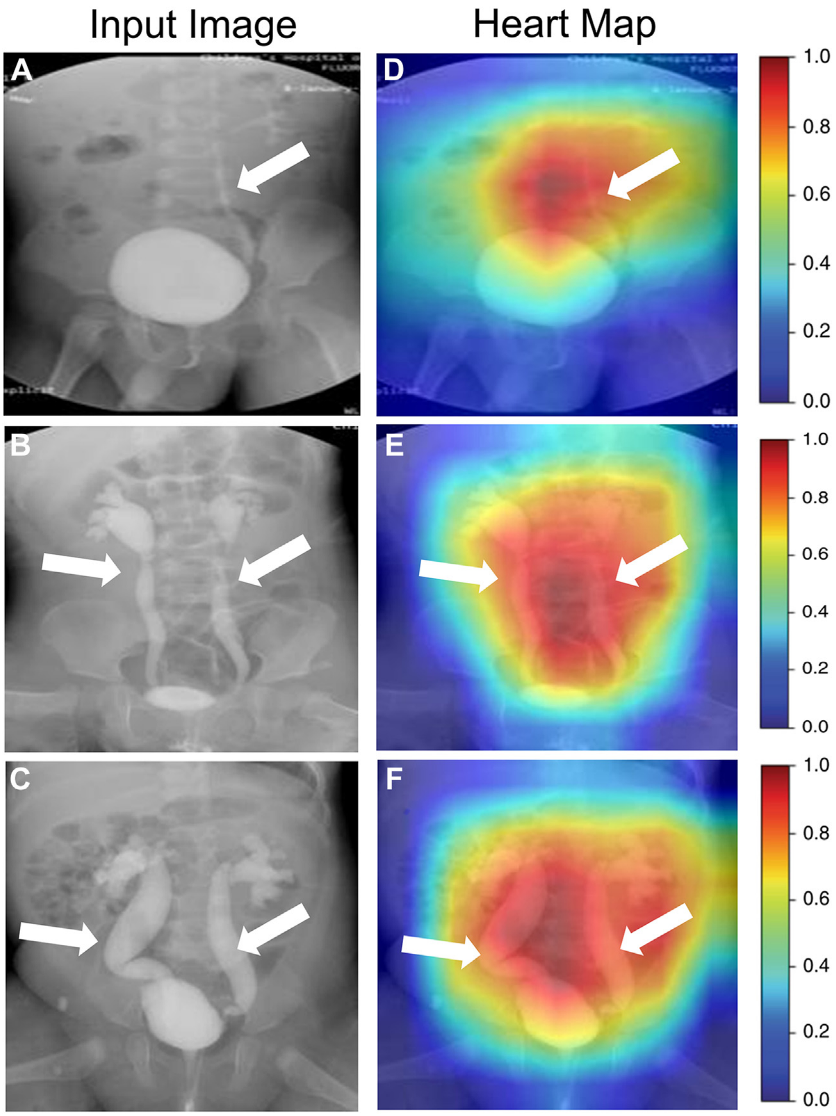

Fig. 7: Patient instances for the visualization and application of the Deep-VCUG model. A, B, C, is the VCUG image, and, D, E, F, is the heat map of the ureter (red indicating higher activation, blue indicating lower activation). The heat maps show that the neural network focused on the region of the ureter for its assessment (arrows).

Previous studies found the inter-rater reliabilities of the VUR grade are very low, with an agreement as low as 60%.[5] In order to improve accuracy of VUR grade, several quantitative metrics have been reported to correlate with VUR, including the distal ureteral diameter ratio (UDR),[11] vesicoureteral reflux Index,[12] sectional area ratio (SAR),[13] renal pelvis dilatation (RPD).[14] Adree Khondker et al[6] developed a machine learning-based model, qVUR, that enables users to determine many of these metrics (such as the ureter-opelvic junction (UPJ) width, ureterovesical junction (UVJ) width, maximum ureter width, and tortuosity of the ureter) from an uploaded image of a VCUG and utilize a random forest classifier was trained to distinguish between low and high grade VUR. Then, in a large pediatric population from multiple institutions, Adree Khondker et al[5] further demonstrate that qVUR could improve reliability for vesicoureteral reflux grading using ureteral tortuosity and dilatation on voiding cystourethrograms. However, this can only identify grade 2 to grade 5 of VUR and these features manually marked is a tedious and time-consuming task. One of the main advantages of CNN is that the features used by the model are learned automatically from the data, rather than being hand-selected by the researcher. Therefore, we developed a multitask learning model, Deep-VCUG, to automatically grade VUR based on a single VCUG image without marking these features.

Our study had several advances compared with the previous study. Firstly, we combined CNNs and ensemble learning to develop a Deep-VCUG model for the grade of VUR from VCUG image to achieve grading reliabilities. Previous, Yesim et al[8] first proposed a hybrid-based Minimum Redundancy Maximum Relevance (mRMR) using the CNN model for the grading of VUR on VCUG images. Of note, use a single CNN for classification, which has limited feature recognition ability. The ensemble learning method combines multiple models to make joint decisions, prediction accuracy will be improved than a single CNN.[15] The voting-ensemble method is a type of computational ensemble method for combining predictions from various sub-models. So, we developed a voting ensemble method that used majority voting to integrate the prediction results of the five individual models, each of which was built on the optimal feature combinations. The sub-model prediction results are statistically compared and analyzed, we found that Deep-VUCG with the highest prediction accuracy than clinicians and five other CNNs in the internal and external test sets.

Second VUR is classified into unilateral or bilateral according to the affected side.[16] Bilateral VUR is often associated with secondary reflux, such as neurogenic bladder or posterior urethral valves. Previous studies have focused on unilateral VUR grading and the grading of bilateral reflux requires manually split sides. However, the evaluation of each side is a simple and objective problem for the clinician. If the DL models could directly output the reflux grading results for each side for bilateral VUR, it could enhance its clinical applicability.

Third, for unilateral and bilateral VUR grade, the Deep-VCUG model demonstrated a higher classification performance compared to two senior clinicians and two junior clinicians in the internal and external test sets. The senior clinicians outperformed the two junior clinicians in the external test sets. We have combined the grading of clinician experts with that of the deep learning model. With such combined, the AUC, sensitivity, specificity, accuracy, and agreement of clinicians were improved. These findings suggest that Deep-VCUG have good performance and under the Deep -VCUG model assist could improve the performance of clinician in clinical.

Fourth, we employed class activation mapping (CAM) to visualize the internal features of the neural network. CAM can help us understand whether the CNN focuses on the suitable area by generating a rough heatmap. In our study, Deep-VCUG focuses on the ureter area indicating that the model learned to assess the correct features instead of learning image correlations.

There are some limitations to this study. First, ground truth was obtained from expert readings. Although expert reading is the best standard of reference for many applications, they might contain variability. Second, the DL model only used a single picture which taken at the peak of the filling phase for its assessment, even though in clinical practice additional phase images for VUR evaluation, such as quiescent period, filling period, and urination period during VCUG. The next task is to combine these additional phase images for multimodal. Third, the DL models with "black box" problems. We visualize and interpret our model using Grad-CAM to clarify the focus of model, alleviating the "black-box" problem of DL models. Fourth, there are additional parameters are not controlled when using a single image of the entire VCUG, such as overlying bowel gas, medical devices, and patient positioning. Five, our study does not include prospective other important clinical data, such as age, gender, urinary tract infection occurrence, and renal scarring. This is the next important area of our future study.

In conclusion, the Deep-VCUG model was able to predict unilateral or bilateral VUR grading with highly objective, reliability and would improve the accuracy and agreement of the clinical with the help of this model. Therefore, use of this model is practical and feasible and has broad application prospects in clinical practice.

Contributors

LZC, ZLT, ZYW, WJT, LKZ, and MW were responsible for concept and design. LZC, ZLT, ZYW, WJT, LKZ, and MW were literature research. ZYW and ZLT provided statistical analysis. All authors were involved in drafting and technical support in deep learning methods. LZC, ZLT, ZYW, WJT, XR, JHF, GBW, HC, JRC, YHD, LKZ, and MW were responsible for acquisition, Processing, or interpretation of data. LZC, ZYW, WJT, LKZ, and WM were involved in drafting the manuscript. All authors had full access to all the data and verified the underlying data. All authors were involved in reviewing the manuscript and approved the final manuscript for submission.

Data sharing statement

Due to the privacy of patients, the data related to patients cannot be available for public access but can be obtained from the corresponding author on reasonable request approved by the institutional review board of all enrolled centers.

Declaration of interests

All authors declare no competing interests.

Acknowledgements

Supported by Natural Science Foundation of China (NSFC82170793), "Fuqing Scholar" Student Scientific Research Program of Shanghai Medical College, Fudan University (FQXZ202113B), and Program of Greater Bay Area Institute of Precision Medicine (Guangzhou) (IPM2021C003).

Prof. Dr. Guangbing Wang (**G.B.M.**) from Puyang People's Hospital; Dr. Han Chu (**H.C.**) from The Children's Hospital of Guangxi Zhuang autonomous region, Dr. Jiarong Chen (**J.R.C.**) from The Affiliated Hospital of Qingdao University, and Dr. Yuhe Duan (**Y.H.D.**) from Anhui Provincial Children's Hospital for providing data as external testing set to support this study.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102466.

References

1 Peters CA, Skoog SJ, Arant BS, et al. Summary of the AUA guideline on management of primary vesicoureteral reflux in children. *J Urol*. 2010;184(3):1134–1144.

2 Tekgül S, Riedmiller H, Hoebeke P, et al. EAU guidelines on vesicoureteral reflux in children. *Eur Urol*. 2012;62(3):534–542.

3 Schaeffer AJ, Greenfield SP, Ivanova A, et al. Reliability of grading of vesicoureteral reflux and other findings on voiding cystourethrography. *J Pediatr Urol*. 2017;13(2):192–198.

4 Metcalfe CB, Macneily AE, Afshar K. Reliability assessment of international grading system for vesicoureteral reflux. *J Urol*. 2012;188(4 Suppl):1490–1492.

5 Khondker A, Kwong JCC, Yadav P, et al. Multi-institutional validation of improved vesicoureteral reflux assessment with simple and machine learning approaches. *J Urol*. 2022;208(6):1314–1322.

6 Khondker A, Kwong JCC, Rickard M, et al. A machine learning-based approach for quantitative grading of vesicoureteral reflux from voiding cystourethrograms: methods and proof of concept. *J Pediatr Urol*. 2022;18(1):78.e1–78.e7.

7 Eroglu Y, Yildirim K, Çinar A, Yildirim M. Diagnosis and grading of vesicoureteral reflux on voiding cystourethrography images in children using a deep hybrid model. *Comput Methods Progr Biomed*. 2021;210:106369.

8 Kwong JCC, McLoughlin LC, Haider M, et al. Standardized reporting of machine learning applications in urology: the stream-uro framework. *Eur Urol Focus*. 2021;7(4):672–682.

9 Medical versus surgical treatment of primary vesicoureteral reflux: report of the International Reflux Study Committee. *Pediatrics*. 1981;67(3):392–400.

10 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2020;128(2):336–359.

11 Cooper CS, Alexander SE, Kieran K, Storm DW. Utility of the distal ureteral diameter on VCUG for grading VUR. *J Pediatr Urol*. 2015;11(4):183.e181–183.e186.

12 Garcia-Roig M, Ridley DE, McCracken C, Arlen AM, Cooper CS, Kirsch AJ. Vesicoureteral reflux index: predicting primary vesicoureteral reflux resolution in children diagnosed after age 24 months. *J Urol*. 2017;197(4):1150–1157.

13 Yi H, Cui X, Cai B, Qiu L, Song P, Zhang W. A quantitative grading system of vesicoureteral reflux by contrastenhanced voiding urosonography. *Med Ultrason*. 2020;22(3):287–292.

14 Hothi DK, Wade AS, Gilbert R, Winyard PJ. Mild fetal renal pelvis dilatation: much ado about nothing? *Clin J Am Soc Nephrol*. 2009;4(1):168–177.

15 Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J Appl Stat*. 2018;45(15):2800–2818.

16 Al Qahtani W, Sarhan O, Al Otay A, El Helaly A, Al Kawai F. Primary bilateral high-grade vesicoureteral reflux in children: management perspective. *Cureus*. 2020;12(12):e12266.