Data and Metadata. 2025; 4:460
doi: 10.56294/dm2025460

ORIGINAL

Automated quantification of vesicoureteral reflux using machine learning with advancing diagnostic precision

Cuantificación automatizada del reflujo vesicoureteral mediante aprendizaje automático para mejorar la precisión diagnóstica

Muhyeeddin Alqaraleh[1], Mohammad Al-Batah[2], Mowafaq Salem Alzboon[2], Esra Alzaghoul[3]

[1]Zarqa University, Faculty of Information Technology. Zarqa, Jordan.
[2]Jadara University, Faculty of Information Technology. Irbid, Jordan.
[3]The University of Jordan, Computer Information Systems Department. Amman, Jordan.

Cite as: Al-batah M, Al-Batah M, Salem Alzboon M, Alzaghoul E. Automated Quantification of Vesicoureteral Reflux using Machine Learning with Advancing Diagnostic Precision. Data and Metadata. 2025; 4:460. https://doi.org/10.56294/dm2025460

Submitted: 20-02-2024 **Revised:** 19-06-2024 **Accepted:** 31-10-2024 **Published:** 01-01-2025

Editor: Adrián Alejandro Vitón-Castillo

Corresponding author: Muhyeeddin Alqaraleh

ABSTRACT

This article uses machine learning to quantify vesicoureteral reflux (VUR). VCUGs in pediatric urology are used to diagnose VUR. The goal is to increase diagnostic precision. Various machine learning models categorize VUR grades (Grade 1 to Grade 5) and are evaluated using performance metrics and confusion matrices. Study datasets come from internet repositories with repository names and accession numbers. Machine learning models performed well across several measures. KNN, Random Forest, AdaBoost, and CN2 Rule Induction consistently scored 100 % in AUC, CA, F1-score, precision, recall, MCC, and specificity. These models classified grades well individually and collectively. In contrast, the Constant model performed poorly across all criteria, suggesting its inability to categorize VUR grades reliably. With the most excellent average performance ratings, the CN2 Rule Induction model excelled at grade categorization. Confusion matrices demonstrate that machine learning models predict VUR grades. The large diagonal numbers of the matrices show that the models are regularly predicted effectively. However, the Constant model's constant Grade 5 forecast reduced its differentiation. This study shows that most machine learning methods automate VUR measurement. The findings aid objective pediatric urology grading and radiographic evaluation. The CN2 Rule Induction model accurately classifies VUR grades. Machine learning-based diagnostic techniques may increase diagnostic precision, clinical decision-making, and patient outcomes.

Keywords: Vesicoureteral Reflux; Voiding Cystourethrogram; Machine Learning; Objective Grading; Radiographic Evaluation; Pediatric Urology.

RESUMEN

Este artículo utiliza aprendizaje automático para cuantificar el reflujo vesicoureteral (RVU). Las cistouretrografías miccionales (VCUG) en urología pediátrica se emplean para diagnosticar el RVU, con el objetivo de aumentar la precisión diagnóstica. Se evaluaron diversos modelos de aprendizaje automático para clasificar los grados de RVU (Grado 1 a Grado 5) utilizando métricas de rendimiento y matrices de confusión. Los conjuntos de datos del estudio provienen de repositorios en línea, con los nombres de los repositorios y los números de acceso correspondientes. Los modelos de aprendizaje automático obtuvieron buenos resultados en varias métricas. KNN, Random Forest, AdaBoost y CN2 Rule Induction lograron consistentemente un 100 % en AUC, CA, F1-score, precisión, recall, MCC y especificidad. Estos modelos clasificaron tanto individual como colectivamente. En contraste, el modelo constante tuvo un rendimiento deficiente en todos los criterios, lo que sugiere su incapacidad para categorizar de manera fiable los grados de RVU. Con las mejores puntuaciones promedio de rendimiento, el modelo CN2 Rule Induction destacó en la categorización de grados. Las matrices de confusión demuestran que los modelos de aprendizaje automático predicen con éxito los grados de RVU. Los altos números diagonales de las matrices muestran que los modelos realizan predicciones acertadas regularmente. Sin embargo, el modelo constante, al predecir siempre el Grado 5, redujo su capacidad de diferenciación. Este estudio demuestra que la mayoría de los métodos de aprendizaje automático automatizan la medición del RVU. Los hallazgos contribuyen a una evaluación objetiva en la urología pediátrica y la evaluación radiográfica. El modelo CN2 Rule Induction clasifica con precisión los grados de RVU. Las técnicas diagnósticas basadas en aprendizaje automático pueden aumentar la precisión diagnóstica, la toma de decisiones clínicas y los resultados de los pacientes.

Palabras clave: Reflujo Vesicoureteral; Cistouretrografía Miccional; Aprendizaje Automático; Clasificación Objetiva; Evaluación Radiográfica; Urología Pediátrica.

INTRODUCTION

AI revolutionizes healthcare and patient care. AI can analyze massive volumes of data, find patterns, and provide significant insights that help in accurate diagnosis, treatment planning, and decision-making in medicine.(1) AI algorithms can swiftly process and interpret medical images, including X-rays, MRIs, and CT scans, helping radiologists spot problems and improve diagnosis accuracy. AI-powered prediction models can also assist doctors in identifying high-risk individuals for particular diseases, enabling early interventions and individualized treatment strategies. AI-driven chatbots and virtual assistants provide 24/7 assistance, improving healthcare information access and lowering physician workload. AI offers medical personnel better tools and skills, improving patient outcomes and efficiency and possibly revolutionizing healthcare delivery.(2)

Vesicoureteral reflux (VUR) is a medical disorder in which urine flows backward into the ureter or kidneys. VUR occurs in 30 % of pediatric urinary tract infection (UTI) cases and can cause renal scarring. VCUG is the standard VUR diagnosis and grading procedure. Current VUR grading is subjective and typically has inter-rater disagreement of up to 60 %, even using the international rating method. Due to this inconsistency, VUR grading should be more objective and uniform. Machine learning may help solve this problem. Healthcare diagnoses and treatment decisions have improved because of AI and machine learning. AI-based systems can improve VUR grading accuracy and efficiency using vast datasets and powerful algorithms.(3)

Machine learning has been studied in medical imaging analysis. Baray et al. developed a deep learning-based penile curvature measurement approach that was accurate in model and patient photos. They also studied VUR grading using quantitative characteristics from voiding cystourethrograms (VCUG) and machine learning. Both investigations demonstrated promising outcomes. However, they were restricted in picture analysis and intermediate VUR grade discrimination.(4) This work addresses past research constraints and develops a machine learning-based VUR grading method utilizing VCUG pictures. We evaluate voiding cystourethrogram pictures to determine ureter and renal pelvis size and shape. These extracted characteristics feed machine-learning algorithms that predict VUR severity. We add elements to better capture ureter tortuosity and distinguish VUR grades.(5)

This work uses machine learning to improve VUR grading accuracy and efficiency. By creating a solid and objective grading system, we want to help doctors stratify patients, manage drugs, and improve VUR care. We expect to outperform subjective methods using robust machine learning classifiers and a wide range of features. In this work, we used VCUG pictures with VUR grades to train and test VUR grading machine learning models. Voiding cystourethrogram pictures were investigated to identify characteristics that might help classify VUR severity.(6)

Various image processing methods extracted ureter size, renal pelvis form, and tortuosity. These features were chosen for their ability to distinguish VUR classes. This study also offered additional characteristics to capture ureter tortuosity and distinguish intermediate VUR grades. Many classifiers were tested to train machine learning models. SVM, random forests, and CNN were among these classifiers. Using extracted features as input, the models learned the link between picture attributes and VUR severity.(7)

The machine learning-based VUR grading system was assessed using accuracy, precision, recall, and F1 score. Cross-validation ensured model robustness and generalizability. The experiments showed that machine learning-based VUR grading works. The models classified VUR grades more accurately than subjective methods. Additional features to capture ureter tortuosity and discriminate VUR grades improved model accuracy.(8) This study has significant clinical implications. A reliable and objective VUR grading system can enhance patient care and therapy. A trustworthy technique can help clinicians stratify patients by VUR severity for appropriate therapies and medication management. The machine learning-based technique may also discover VUR grading problems, improving diagnostic quality control.(9)

This work is a significant step toward objective VUR measurement using machine learning. However, it has limits. This study should employ a bigger, more diversified group of VCUG pictures to guarantee model generalizability. The machine learning-based technique has to be tested in clinical situations and compared to other grading systems.(10) Finally, VUR must be graded accurately and objectively for successful management and therapy. Current subjective VUR grading methods show considerable inter-rater discrepancy, highlighting the need for a more uniform and standardized methodology. This work proposes a machine learning-based VUR grading system utilizing VCUG pictures to improve accuracy and efficiency over subjective methods.(11)

VCUG pictures and related characteristics were used to train machine learning algorithms to predict VUR severity. Adding characteristics to capture ureter tortuosity enhanced VUR grade distinction. The accurate models might improve VUR management and treatment decision-making.(12) This study is promising, but further research is needed to confirm the machine learning-based strategy on more extensive and varied datasets. The VUR grading system should also be tested in clinical settings and compared to other methods. Machine learning in VUR grading might improve patient outcomes and standardize diagnosis. Clinicians may effectively stratify and manage patients using modern algorithms and image analysis to objectively and accurately determine VUR severity.(13)

Problem Statement

The radiographic vesicoureteral reflux (VUR) assessment is currently based on subjective criteria and is susceptible to variability, resulting in inconsistent diagnoses and treatment suggestions.(14) Subjectivity presents challenges for clinicians and radiologists when evaluating VUR severity using voiding cystourethrogram (VCUG) images. Hence, creating a precise and unbiased grading system for VUR is necessary to enhance diagnostic accuracy and treatment results.(15)

Article Objectives

The primary objective of this article is to create a supervised machine-learning model to grade VUR objectively using VCUG images. To tackle the subjectivity and variability in VUR grading using machine learning algorithms.(16) To assess the effectiveness of various machine learning models in accurately predicting the severity of VUR across different grade levels. To identify the key features and patterns in VCUG images that help accurately classify VUR grades. To evaluate if the deformed renal calyces pattern can predict high-grade VUR. To show that machine learning can improve objectivity and accuracy in VUR grading, resulting in reliable diagnostic assessments.(17)

Contribution of the Article

The article enhances radiology and healthcare AI research by utilizing supervised machine learning techniques to improve the grading of VUR. The main contributions of the article are, Initially, the creation of machine learning models: The article discusses the creation and assessment of six machine learning models for grading VUR using VCUG images: Logistic Regression, Tree, Gradient Boosting, Neural Network, and Stochastic Gradient Descent. The article shows that machine learning models effectively predict VUR severity across various grade levels without false positives or negatives.(18) This discovery emphasizes the capability of machine learning to address subjectivity and variability in VUR grading. Thirdly, Identification of Key Features: The article highlights that deformed renal calyces are a strong indicator of severe vesicoureteral reflux. This discovery offers valuable information for clinicians and radiologists to improve the diagnosis and treatment of VUR. The article recognizes the necessity for additional enhancement of machine learning processes, investigation of novel features, and enlargement of the dataset to improve the precision and applicability of the models.(19) This discovery sets the stage for upcoming studies and enhancements in VUR grading through machine learning techniques. 5. Implications for Healthcare: The article highlights how machine learning-based grading systems can decrease subjectivity and unpredictability in VUR assessments. The proposed method can enhance objectivity and accuracy, resulting in more reliable VUR grading and improving patient diagnoses and treatment recommendations. The article enhances the field of VUR grading by using machine learning techniques to improve accuracy, objectivity, and efficiency in evaluating VCUG pictures.(20)

Article Organization

Section 2 summarizes the literature on VUR prediction and classification models. Section 3 outlines the proposed methodology, which includes a description of the dataset and the machine learning model. Section 4 demonstrates the implementation and its outcomes.(21) Section 5 contains the discussions. The paper is concluded in section 6.

RELATED WORK

GERD prevalence varies globally and affects community health. Chronic GERD causes several illnesses. The incidence rate, risk factors, and symptoms of GERD must be determined to improve healthcare and management. Machine learning was used to gather data for this study. Data was categorized using Artificial Neural Network principles, and VOSviewer software produced the results as a network. Artificial intelligence and machine learning show that Asian GERD rates are growing. Pakistani reports say GERD is prevalent in various locations. In Pakistan, oily food, late dinners, sedentary lifestyles, and lack of understanding about illness diagnosis and GERD treatment are risk factors. The illness involves acid reflux and esophageal inflammation, according to our results. This study will examine risk variables and symptom incidence to improve GERD diagnosis. Healthcare professionals might easily monitor GERD patients to reduce the illness burden of GERD and associated diseases. Identifying geographical variances and evaluating comparative data can help identify disease hotspots that require additional disease management and control.(22)

This study developed and validated prediction models for persistent chronic cough (PCC) in chronic cough patients. The study was retrospective and cohort-based. Approaches A specialist cohort of specialists diagnosed patients aged 18 to 85 and an event cohort of patients with at least three coughs were found from 2011 to 2016. A cough diagnosis, prescription, or clinical record mention is a cough occurrence. With almost 400 features and two machine-learning methods, model training and validation were done. We did sensitivity analyses. PCC was diagnosed by CC or two cough occurrences in years two and three after the index date for the specialist cohort or three for the event cohort. Outcomes: 8581 patients met the expert cohort criteria, and 52,010 met the event cohort criteria, with mean ages of 60,0 and 55,5 years. 38,2 % of specialists and 12,4 % of event cohort patients got PCC. Utilization-based models examined baseline CC or respiratory illness healthcare consumption. However, diagnosis-based models included age, asthma, pulmonary fibrosis, obstructive lung illness, gastro-oesophageal reflux, hypertension, and bronchiectasis. The succinct models with five to seven predictors had reasonable accuracy, with the area under the curve values of 0,74 to 0,76 for utilization-based models and 0,71 for diagnosis-based models. Conclusions Our risk prediction algorithms can identify high-risk PCC patients during clinical testing/evaluation to improve decision-making.(23)

Voice analysis's cutting-edge topic is automated speech disorder evaluation. A recent study suggests eating disorders may affect voice qualities. This study evaluated how obesity and GERD affect voice, with obesity being a risk factor. We discussed how diseases interact with consistent features. Vowel phonation and sentence repetition were tested on 92 people. Healthy controls, obese patients, and obese GERD patients were studied. The Naive Bayes and Support Vector Machine models extracted binary classification features well. They detected GERD and obesity with 0,86 and 0,82 accuracy on the validation set. Performance dropped on the test set, indicating no overfitting. Vowel phonation was less successful than sentence repetition in tasks and characteristics. Mel Frequency Cepstral Coefficients, Perceptual Linear Prediction Coefficients, Bark Band Energy Coefficients, and noise measurements are most important for the application.(24)

Drying, grinding, and refluxing pepper samples with high-quality ethanol to measure piperine content is laborious. This method is efficient but time- and resource-intensive, limiting its ability to address variances. A pressing need is to find faster and more accurate machine learning methods for measuring and predicting piperine content. Fluorescence imaging and artificial neural network (ANN) models are tested to improve Javanese long pepper piperine measurement accuracy. Our proposal analyzes Javanese long pepper using UV-induced fluorescence imaging and machine learning. UV LEDs at 365 nm created fluorescence, with varied hues representing piperine concentration. Based on fluorescence picture color texture properties, an artificial neural network model predicted piperine content with an R2 value of 0,88025. Ten selected features and One-R were used in the model. The ultimate artificial neural network (ANN) has a testing R2 of 0,8943 and MSE of 0,0875 utilizing 'trainees' learning, 'tansig' activation, 0,1 learning rate, and 10-40-10 nodes. LED fluorescence helps machine learning piperine prediction. The effectiveness of piperine content measurement is increased in this work.(25)

There are conflicting reports on the parameters needed for successfully correcting vesicoureteral reflux with dextranomer/hyaluronic acid copolymer. We conducted a logistic regression analysis to assess the impact of injected volume while controlling for other variables that may be linked to the success of dextranomer/hyaluronic acid copolymer injection. Methodology: From July 2003 to June 2006, 126 consecutive patients (34 males and 92 females) with an average age of 6,5 ± 3,7 years and primary vesicoureteral reflux (196 refluxing ureters) received injections for febrile urinary tract infections. The complete resolution of reflux determined success. Analyzed factors were age, gender, laterality, preoperative vesicoureteral reflux grade, surgeon experience, dextranomer/hyaluronic acid copolymer volume, time to surgery from initial presentation, and preoperative treatment for lower urinary tract symptoms. The results showed that Vesicoureteral reflux grades ranged from I to V in 7 (3,5 %), 53 (27 %), and 91 (46,4 %).(26)

This study examines depression and particulate matter as preterm birth risk factors using machine learning and demographic data. The retrospective cohort study used Korea National Health Insurance Service claims for

405,586 25–40-year-old women who gave birth for the first time after a singleton pregnancy between 2015 and 2017. From 2015 to 2017, 90 independent variables covered demographic, socioeconomic, environmental, health, and obstetric aspects, with preterm birth as the dependent variable. Depression and particulate particles were found to cause preterm birth using random forest variable importance. According to random forest variable importance, socioeconomic status, age, proton pump inhibitor, benzodiazepine, a tricyclic antidepressant, sleeping pills, progesterone, gastroesophageal reflux disease (GERD) for 2002–2014, particulate matter for January–December 2014, region, myoma uteri, diabetes for 2013–2014, and depression for 2011–2014 are the top 40 determinants of preterm birth in 2015–2017. Particulate particles and depression are highly associated with preterm birth. Effective prenatal care includes rigorous particle intervention, proactive counseling, and medication for common depression symptoms that expecting mothers ignore.(27)

This work predicts crude oil cutbacks in the initial refining stage using rough set theory (RST) and an adaptive neuro-fuzzy inference system (ANFIS) soft sensor model to increase oil refinery efficiency. The ANFIS model's fuzzy rule sets and decision table 2 properties were simplified using the RST. Optimal discretization algorithms were employed for continuous data. This predicts Reid Vapor Pressure (RVP) and API gravity, influencing light naphtha cut quality. Real-time process data from the Al Doura oil refinery is analyzed to process crude oil from two sources better. The response variables show the cascade controller's feedback value at the splitter's top in the crude distillation unit's rectifying part. A controller controls reflux liquid flow to the splitter head. The adaptive soft sensor concept created a steady-state control system with an integrated virtual sensor that matched laboratory tests. A predictive control system using a cascade ANFIS controller and a soft sensor model maintains distillate purity within an oil refinery's quality control range. The ANFIS-based cascade control has no over/undershoots and improves rising time by 26,65 % and settling time by 84,63 % over the PID-based control. Additionally, prediction and control model results are compared to other machine learning approaches.(28)

Determine the number of STING-based endoscopic injection (EI) operations needed for best performance without simulated training. Methodology: Two pediatric urology fellows' EI procedures were investigated. Patients without primary vesicoureteral reflux, endoscopic injection, ureteroneocystostomy, lower urinary tract dysfunction, or duplicated ureters were excluded from the research. Following O'Donnell and Puri's methodology, the researchers used dextranomer hyaluronate and STING. Numerous statistical trials determined group size. One combination with 35 EI surgeries and three combinations with 12, 24, and 36 patients showed statistically significant changes. Thus, 12 patient groups were developed. The first completed 54 EIs, the second 51. Therefore, the first colleague had three groups of 18 EI procedures, whereas the second had 17 in each group. The research has 72 participants and 105 ureter units. The three categories had 35 procedures when both individuals' data were pooled. The first person's success rates were 38,3 %, 66,6 %, and 83,3 % in the first, second, and third groups, respectively, with p = 0,02. Success rates for the second person were 41,2 %, 64,7 %, and 82,3 %, with p = 0,045. Both colleagues' success rates rose similarly. CONCLUSIONS: EI may be successful after 20 sessions and reasonable after 35-40.(29)

Contrast-enhanced voiding ultrasonography (ceVUS) was used to identify vesicoureteral reflux (VUR) and assess intrarenal reflux (IRR). The study attempted to identify IRR VUR patients and their kidney locations. Methodology: Seventy patients with vesicoureteral reflux (VUR) and 103 uretero-renal units (URUs) had contrast-enhanced voiding ultrasonography (ceVUS) for recurrent febrile urinary tract infections (UTI) or first febrile UTI with renal ultrasonography abnormalities. Patients were examined using the GE Logiq S8 ultrasound machine and second-generation ultrasound contrast. Intrarenal reflux (IRR) was found in 51 (49,51 %) of 103 instances of vesicoureteral reflux (VUR), regardless of VUR severity ($p < 0,0001$). IRR diagnostic median age was five months (IQR, 3-14,3), compared to 15,5 months (IQR, 5-41,5) for patients without IRR ($p = 0,0069$). IRR incidence was highest in the superior pole (80 %), followed by inferior (62,7 %), middle (37 %), and all segments (27 %), $p < 0,0001$. Conclusion: This study found early clinical signs in IRR-associated VUR patients. IRRs followed the normal distribution of composed papillae types II and III and increased with VUR severity. Future clinical studies may underline the need to integrate IRR in VUR classification.(30)

Vesicoureteral reflux patients' clinical development and treatment choices are often assessed using radiographic grading of voiding cystourethrogram (VCUG) images. Since image-based VUR grading is subjective, we constructed a supervised machine-learning algorithm to rate VCUG data objectively. This study used 113 public VCUG images. Four pediatric radiologists and three pediatric urologists assessed VUR severity in each scan. Severity was 4-5, whereas low severity was 1-3. For each photograph, the grade most expert assessors indicated was the ground truth. Nine features were retrieved from each VCUG picture, and six machine learning models were trained, validated, and tested using 'leave-one-out' cross-validation. All attributes were analyzed, and the best were used to train models. The F1-score is a prominent machine learning model accuracy statistic. The support vector machine (SVM) and multilayer perceptron (MLP) classifiers have excellent accuracy with F1 scores of 90,27 % and 91,14 %, respectively, using the most critical VCUG image attributes. When all features were used, SVM and MLP F1 scores were 89,37 % and 90,27 %, respectively. According to the findings, an aberrant renal calyces pattern indicates severe VUR. Future machine-learning techniques can enhance VUR objective grading.[31]

High-resolution manometry (HRM) and esophagography are used to diagnose achalasia. However, the esophageal motility and morphological phenotypes are unknown. Per-oral endoscopic myotomy (POEM) outcomes in new patients are difficult to predict. This multicenter cohort study included 1,824 treatment-naïve achalasia patients. Overall, 1,778 patients got POEM. Based on patients' demographic data such as age, sex, illness duration, BMI, and HRM/esophagography findings, machine learning clustering was utilized to identify achalasia phenotypes. Chronic symptoms with an Eckardt score of 3 or higher and reflux esophagitis graded A to D after ML models predicted POEM.ML classified achalasia into three phenotypes: type I with a dilated esophagus (n=676; 37,0 %); type II (n=203; 11,1 %); and late-onset type I-III (n=619, 33,9 %). Phenotypes 1 and 2 of Types I and II achalasia have different clinical symptoms than phenotype 3, suggesting different pathophysiologies within the same HRM diagnosis. The persistent symptom prediction model achieved an AUC of 0,70. Pre-POEM Eckardt scores of 6 or above were the most significant cause of persistent symptoms. POEM reflux esophagitis (RE) AUC was 0,61. distinct types of achalasia, which include esophageal mobility and structure, imply distinct causes. By offering fresh views on treatment resistance factors, machine learning helped create an effective risk categorization model for persistent symptoms.[32]

The stomach, esophagus, duodenum, small intestines, and large intestinal tract are fundamental to human physiology. Many people worldwide suffer gastric dysrhythmia, dyspepsia, unexplained nausea, vomiting, abdominal pain, stomach ulcers, and GERD. Clinical analysis, endoscopy, electrogastrogram, and imaging detect irregularities. A stomach electrogastrogram records electrical impulses that constrict stomach muscles. Electrogastrograms are recorded when the electrode detects stomach muscle electrical impulses. Computers handle electrogastrogram (EGG) signals. Normal stomach muscle electrical activity rises after meals. Stomach muscle or nerve anomalies cause postprandial electrical rhythm abnormalities. This research analyzes average electrogastrograms (EGGs) for bradycardia, dyspepsia, nausea, tachycardia, ulcers, and vomiting. Data is collected before and after meals with the doctor for patients and the general public. The MATLAB genetic method employs Continuous Wavelet Transform (CWT) and the db4 wavelet to depict an EGG signal wave pattern in 3D. The peak signifies the EGG signal cycle, as seen in the picture. Peak count classifies EGG. The Adaptive Resonance Classifier Network (ARCN) classifies EGG signals as normal or abnormal according to attention (μ). This study can help doctors diagnose stomach disorders before invasive procedures. The suggested work has 95,45 % accuracy, 92,45 % sensitivity, and 87,12 % specificity.[33]

Vesicoureteral reflux grading from voiding cystourethrograms is subjective and unreliable. Simple and machine-learning approaches concentrating on ureteral tortuosity and dilatation in voiding cystourethrograms were used to improve vesicoureteral reflux grading reliability. Our institution provided voiding cystourethrograms for training and five sets for validation. For each voiding cystourethrogram, 5-7 raters established a consensus grade for vesicoureteral reflux and measured rating reliability between and among raters. Each voiding cystourethrogram was evaluated for ureteral tortuosity proximal, distal, and maximal dilatation. Labels were then applied to the four characteristics. A machine learning model, qVUR, predicted vesicoureteral reflux grade using particular factors. AUROC study assessed model performance. Voiding cystourethrograms collected 1,492 kidneys and ureters, resulting in 8,230 assessments. Vesicoureteral reflux grading had 0,44 internal consistency, 0,71 median agreement, and low rater consistency. More values for each feature indicated more vesicoureteral reflux. The qVUR performed consistently across all external datasets with an accuracy of 0,62 (AUROC=0,84). The strategy improved vesicoureteral reflux grade reliability 3,6 times over standard methods (P <.001). We show that using machine learning to grade vesicoureteral reflux from many institutions in a large pediatric cohort is more reliable than current methods. The qVUR model is generalizable and robust, with doctor-like precision. The predictive power of quantitative indicators is needed for further study.[34]

Antibiotic prophylaxis regularly cuts urinary tract infections in half in children with vesicoureteral reflux. However, focused antibiotic prophylaxis may help certain groups. Our goal was to develop a machine-learning strategy to discover these categories. We used RIVUR data randomly split into training and testing sets 4:1. Recurrent urinary tract infections were predicted using two models with and without antibiotic therapy. The test set verified recurrent urinary tract infections and antibiotic prophylaxis. Each model predicted recurrent urinary tract infections.

Continuous antibiotic prophylaxis was administered at varying degrees of lowered urinary tract infection risk to evaluate its efficacy. The research included 607 patients, 558 girls and 49 males, with a median age of 12 months. Vesicoureteral reflux grade, serum creatinine, race, gender, history of UTI symptoms (fever/dysuria), and weight percentiles were evaluated. The AUC for the recurrent urinary tract infection prediction model with continuous antibiotic prophylaxis against placebo is 0,82 (95 % CI 0,74-0,87). To reduce the incidence of recurrent urinary tract infections by 10 %, 40 % of vesicoureteral reflux patients should get continuous antibiotic prophylaxis. Out of 121 test subjects, 51 got continuous antibiotic prophylaxis as advised by the model (if recurrent urinary tract infection risk decrease was more than 10 %). This group had fewer recurrent urinary tract infections than those whose continuous antibiotic prophylaxis assignment differed from the model

(7,5 % vs 19,4 %, p=0,037). Our prediction algorithm can identify vesicoureteral reflux patients most benefit from ongoing antibiotic prophylaxis. This method targets and individualizes continual antibiotic prophylaxis, enhancing its efficacy and reducing wasteful usage in potentially unneeded patients.(35)

Most clinicians use the Los Angeles classification to diagnose and treat GERD, a common gastrointestinal illness. Advanced artificial intelligence allows deep learning models to help physicians diagnose. This work develops a two-stage endoscopic classification approach for GERD using deep learning and machine learning. It uses transfer learning on the target dataset to improve picture feature extraction and machine learning methods to optimize classification. The experimental results demonstrate that this study's GerdNet-RF model outperforms others. Increase test accuracy from 78,8 % ± 8,5 % to 92,5 % ± 2,1 %. Improving AI model automated diagnostics would improve patient healthcare.(36)

In vesicoureteral reflux (VUR), urine runs backward from the bladder into the ureters and possibly the kidneys. It is essential to urinary tract infections. VCUG imaging evaluates the severity of vesicoureteral reflux (VUR). The time and type of VUR surgery are debated. Identifying VUR grades regularly and accurately is crucial. This study aims to properly detect and characterize VUR in VCUG pictures using a convolutional neural network (CNN). The goals are to reduce categorization disparities among observers and create an accessible tool for healthcare practitioners.(37)

NSAIDs are recommended worldwide to treat pain and inflammation, especially in osteoarthritis. NSAIDs are known to harm the gut. Patients and doctors must adjust osteoarthritis medication if a stomach ulcer develops, increasing costs and difficulty. Thus, building a stomach ulcer prediction model that precisely represents each person's health is crucial for treatment planning. We constructed a prediction model for NSAID-induced stomach ulcers using South Korean National Health Insurance Service sample cohort data from 2008 to 2013. Our study used machine-learning algorithms to identify new drug and comorbidity risk factors. From 2008 to 2013, 30,808 osteoarthritis patients got NSAIDs. A 2-year follow-up divided the patients into 29,579 without stomach ulcers and 1,229 with them.

A gradient boosting machine (GBM) was the best prediction model out of five machine-learning approaches, with an area under the curve of 0,896 and a 95 % confidence range of 0,883 to 0,909. Loxoprofen, aceclofenac, talniflumate, meloxicam, and dexibuprofen were identified by the Gradient Boosting Machine (GBM) as essential features, along with AURI and gastroesophageal reflux sickness. No dose-response relationship was found for AURI. It was not a significant risk factor despite being first recognized as an essential trait and improving prediction accuracy. We used GBM to predict NSAID-induced stomach ulcers. Quantifying prescription duration and comorbidity severity can correctly indicate patient risk. The prediction model performed well and was easy to comprehend, benefiting doctors and NSAID users.(38)

LPRD is frequent and has numerous symptoms. LPRD has no symptoms or clinical criteria, making diagnosis difficult. A clinical practice needs an objective, reliable test without a clinical gold standard. Approaches: 60 healthy persons and 74 diagnosed LPR patients were characterized using the reflux symptom index, reflux finding score, 24-hour oropharyngeal pH monitoring, and anti-reflux medication response. pH data yielded 72 properties, and stepwise wrapping found the best combination. Semi-supervised learning blended 1552 unlabeled and labeled data for feature selection and model training. Latent class model evaluation with 64 extra validation data and an inadequate clinical reference test assessed the recommended model. A new W score increased LPRD test sensitivity to 82,67 % from 24,09 % and specificity to 80,19 %. W score matches thorough clinical exam. Conclusion: The W score accelerates LPRD diagnosis and aids clinical diagnosis. The W score is a reliable, accurate, and clinically relevant assessment of anti-reflux drug appropriateness. Significance: LPRD patients who use anti-reflux medication more often can avoid the side effects of unnecessary long-term PPI therapy.(39)

Early renal and anatomical factors may signal future progression and the need for further procedures in posterior urethral valve patients. Machine learning (ML) was used to predict clinically meaningful results in these individuals. We included posterior urethral valve (PUV) patients who underwent renal function tests at our institution from 2000 to 2020. Each visit's estimated glomerular filtration rate (eGFR), initial vesicoureteral reflux grade, and presenting renal dysplasia were recorded. The progression of CKD, the start of KRT, and the need for clean-intermittent catheterization were predicted by ML models. After evaluating the model's concordance index (c-index), external validation was done. 103 individuals were studied, with a median follow-up of 5,7 years. 26 patients (25 %) had chronic renal disease, 18 (17 %) needed kidney replacement, and 32 (31 %) received chronic interstitial cystitis treatment. For external validation, 22 patients were chosen. The machine learning model performed better than Cox proportional-hazards regression in predicting chronic kidney disease, kidney replacement therapy, and chronic interstitial cystitis (c-index=0,77; external C-index=0,78). The models are in user-friendly software at https://share.streamlit.io/jcckwong/puvop/main/app.py. Machine learning can predict clinical outcomes of posterior urethral valves. More verification is needed. However, this practical approach can aid decision-making. The Supplementary material includes a higher-resolution Graphical abstract.(40)

This study examines preterm birth (PTB) with dental and gastrointestinal disorders using machine learning and extensive population data. Approaches: Population-based retrospective cohort data from Korea National Health Insurance claims for 124,606 primiparous women aged 25–40 who gave birth in 2017 were used. The study included 186 independent variables, including demographics, socioeconomics, illness, and medication data. A PTB prediction model was created using machine learning. Random forest variable importance was used to identify PTB risk variables and their relationships to dental and gastrointestinal diseases, medication history, and socioeconomic status. The random forest model with oversampled data has an accuracy of 84,03 % and a receiver-operating-characteristic curve area of 84,03–84,04. PTB is strongly associated with socioeconomic status (0,284), age (0,214), and medical problems and drugs, including GERD in different years, progesterone, tricyclic antidepressants, and infertility. If socioeconomic status, 2014 GERD, and 2016 infertility are randomly mixed, the model's accuracy will drop by 28,4 %, 2,6 %, or 1,9 %. In conclusion, machine learning created a viable PTB prediction model. PTB is associated with GERD and infertility. Pregnant women need obstetric and gastrointestinal surveillance.(41)

For two decades, endoscopic injection (EI) has proven a reliable alternative to open surgery for treating children's vesicoureteral reflux (VUR). This study reviews contentious indications, bulking agents, injectable processes, success factors, and situational usage. Minorly intrusive and well-accepted by patients and families, elective intubation has a short learning curve and low morbidity. Reflux resolution rates are 69 %-100 %, equivalent to open reimplantation. Various factors affect success. It is now the primary therapy for high-grade reflux and complex anatomical abnormalities such as duplex kidneys, bladder diverticula, and ectopic ureters. Deflux and Vantris are the primary injectable materials. The first choice is absorbable, more straightforward to inject, and less likely to block, although it may lose efficacy with time. The second alternative is non-absorbable, challenging to inject, and more likely to block, but it may last longer. The main approaches are STING and HIT. Depending on their knowledge, surgeons choose the best material and injection method.(42)

METHOD

This methodology outlines the research design, data collection methods, and data analysis techniques employed in a study focused on quantifying vesicoureteral reflux (VUR) using machine learning. The study utilized a dataset of VCUG images of real-world VUR cases gathered from various public sources. After eliminating poor-quality images for grading, 113 high-quality photos were selected for analysis. The severity of VUR in each image was independently graded by seven professional assessors, including three pediatric urologists and four pediatric radiologists. The chosen methods align with the research objectives of developing an automated model for VUR severity quantification as shown in figure 1. The dataset used in the study is publicly available on Kaggle (https://www.kaggle.com/datasets/saidulkabir/vcug-vur-dataset).(43)

Machine Learning Models

The evaluation was conducted using the VCUG VUR Dataset, consisting of 113 rows and 6 columns. The dataset includes a categorical outcome variable with 5 classes representing the severity of vesicoureteral reflux (VUR). The dataset contains three numeric variables and two text variables as metas.(44)

Multiple machine learning models were utilized to perform the evaluation. The following models were employed: kNN, Random Forest, AdaBoost, CN2 Rule Induction, and Constant. Each model was trained and evaluated using appropriate performance metrics to assess its effectiveness in predicting the VUR severity.(45)

Regarding data availability, the VCUG VUR Dataset used in this study can be accessed through the online repository Kaggle. The dataset can be found at the following URL: [https://www.kaggle.com/datasets/saidulkabir/vcug-vur-dataset]. Researchers and interested individuals can access the dataset from this repository for further analysis and exploration.(46)

In the evaluation process, several machine learning models were employed to assess their performance in predicting the severity of vesicoureteral reflux (VUR) using the VCUG VUR Dataset. The following models were utilized:

kNN (k-Nearest Neighbors)

kNN is a simple yet powerful machine learning algorithm for classification and regression tasks. It classifies a new data point by finding the k nearest neighbors in the training dataset based on a distance metric (e.g., Euclidean distance) and assigning the majority class label among those neighbors. The formula for kNN can be represented as:

For classification: Classify(x) = mode(y_1, y_2, ..., y_k), where y_i is the class label of the i-th nearest neighbor.

For regression: Predict(x) = mean(y_1, y_2, ..., y_k), where y_i is the target value of the i-th nearest neighbor.

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions.

Each decision tree is built using a random subset of features and a random subset of the training data. The projections from individual trees are combined through voting or averaging to obtain the final prediction. The formula for Random Forest involves aggregating the predictions of multiple decision trees:

For classification: Classify(x) = mode(Classify_i(x)), where Classify_i(x) is the class label predicted by the i-th decision tree.

For regression: Predict(x) = mean(Predict_i(x)), where Predict_i(x) is the target value predicted by the i-th decision tree.

AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble learning algorithm that combines multiple weak classifiers to create a robust classifier. It assigns weights to each training example, with misclassified examples receiving higher weights. The weak classifiers are then trained iteratively, with each subsequent weak classifier focusing on the misclassified examples from the previous iteration.(47) The final prediction is obtained by aggregating the predictions of all weak classifiers, weighted by their performance. The formula for AdaBoost can be expressed as:

For classification: Classify(x) = sign(Σ(α_i * Classify_i(x))), where α_i is the weight assigned to the i-th weak classifier.

For regression: Predict(x) = Σ(α_i * Predict_i(x)), where α_i is the weight assigned to the i-th weak regressor.

CN2 Rule Induction

CN2 Rule Induction is a rule-based machine learning algorithm for classification tasks. It generates a set of if-then rules based on the training data, where each Rule consists of a condition and a corresponding class label. The algorithm uses a beam search strategy to grow and prune rules iteratively, selecting the most accurate and concise rules. The formula for CN2 Rule Induction involves generating and evaluating rules:

For classification: Classify(x) = Rule(x), where Rule(x) is the class label assigned based on the matching Rule's condition.

Constant

The Constant model is a simple baseline model that always predicts a constant value, regardless of the input. It is commonly used as a reference point for evaluating the performance of more complex models. The formula for the Constant model is straightforward:

For classification: Classify(x) = c, where c is a constant class label.

For regression: Predict(x) = c, where c is a constant target value.

Each model was likely evaluated based on various performance metrics, such as accuracy, precision, recall, and F1 score, to determine their effectiveness in predicting VUR severity. The evaluation process helps identify the model or combination of models that yield the best predictive performance for the given dataset.(48)

Research Design

The research approach employed in this study is quantitative, as it involves the analysis of numerical data obtained from VCUG images. The study focuses on developing a machine-learning model for automated VUR severity quantification.(49)

Data Collection Methods

The dataset used in this study was obtained from various public sources, including articles, online resources, and Radiopaedia. This approach ensures the inclusion of real-world VUR cases and a diverse range of scenarios. Images with poor quality for grading were excluded from the dataset to ensure the reliability and accuracy of the analysis.(50)

Sample Selection

The sample for this study comprises 113 VCUG images. These images were selected based on their suitability for accurate VUR severity assessment. The sample represents a variety of VUR cases, allowing for a comprehensive analysis.(51)

Data Collection Procedures

Each of the 113 VCUG images was independently evaluated and graded by seven professional assessors. This group included three pediatric urologists and four pediatric radiologists. The assessors utilized their clinical expertise and knowledge to assign severity grades to each image, considering the extent and severity of VUR present.

Data Analysis Techniques

Machine learning techniques were employed to develop a model for automated VUR severity quantification. Specific feature extraction techniques were applied to capture relevant characteristics from the VCUG images. Various machine learning algorithms, such as convolutional neural networks (CNNs), were explored and evaluated for their effectiveness in quantifying VUR severity.

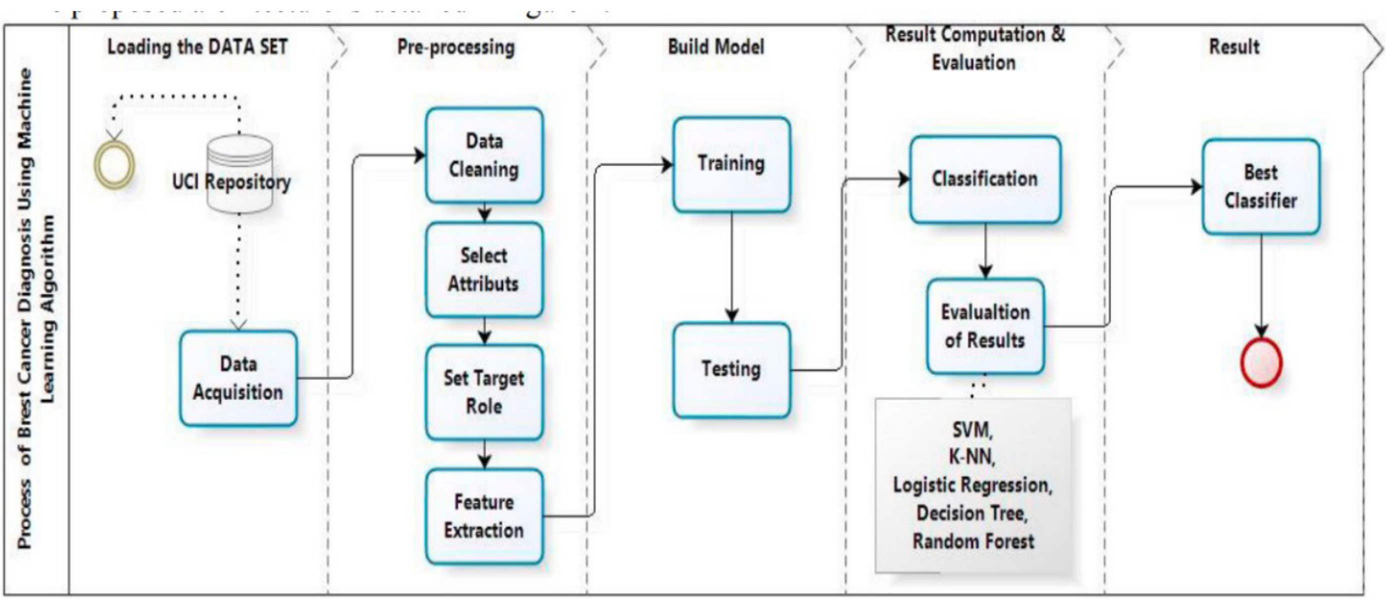

Figure 1. Machine Learning Modelling Processes

RESULTS

Test and Score Analyses

Test and score are fundamental concepts in the realm of evaluation and assessment. A test is a standardized procedure or assessment tool designed to measure a specific construct, such as knowledge, skills, or abilities, usually within a defined domain or subject area. Tests are used in various fields, including education, psychology, and research, to gather data and make informed judgments about individuals or systems. On the other hand, a score represents the numerical or qualitative result obtained from a test, indicating the performance or proficiency level of an individual or the effectiveness of a system. Scores can be expressed as raw scores, percentile ranks, standard scores, or other forms of measurement, depending on the nature of the test and the intended interpretation. The relationship between tests and scores is crucial in evaluating and comparing performances, making decisions, and providing feedback for improvement. Carefully analyzing and interpreting test scores contribute to informed educational, clinical, and organizational decision-making processes as shown in table 1 and figure 2.

Table 1. Test and score analyses for the models KNN, Random Forest, AdaBoost, CN2 Rule Induction, and Constant 1 - 5

Grade	Model	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
Grade 1	KNN	1	1	1	1	1	1	1	0
	Random Forest	1	1	1	1	1	1	1	0,019
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,141
	Constant	0,5	0,963	0	0	0	0	1	0,16
Grade 2	kNN	1	1	1	1	1	1	1	0
	Random Forest	1	0,988	0,963	1	0,929	0,956	1	0,15
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,239
	Constant	0,5	0,825	0	0	0	0	1	0,468
Grade 3	kNN	1	1	1	1	1	1	1	0
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,265
	Random Forest	1	0,988	0,98	0,962	1	0,972	0,982	0,186
	Constant	0,5	0,688	0	0	0	0	1	0,621

Grade 4	kNN	1	1	1	1	1	1	1	0
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,239
	Random Forest	1	0,988	0,968	0,938	1	0,961	0,985	0,15
	Constant	0,5	0,812	0	0	0	0	1	0,483
Grade 5	kNN	1	1	1	1	1	1	1	0
	Random Forest	1	0,988	0,978	1	0,957	0,97	1	0,131
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,278
	Constant	0,5	0,287	0,447	0,287	1	0	0	0,601
Average over classes	kNN	1	1	1	1	1	1	1	0
	AdaBoost	1	1	1	1	1	1	1	0
	CN2 Rule Induction	1	1	1	1	1	1	1	0,657
	Random Forest	1	0,975	0,975	0,976	0,975	0,967	0,991	0,336
	Constant	0,5	0,287	0,128	0,083	0,287	0	0,713	1,47

Test and Score Analyses for Grade 1

In this study, we analyze the performance of different models for the classification of VUR (vesicoureteral reflux) using an online dataset. The table 1 presents performance metrics for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss. The kNN model achieves perfect scores across all metrics, indicating its excellent performance in accurately classifying VUR instances. It attains an AUC of 1, demonstrating discriminative solid ability. The classification accuracy, F1-score, precision, recall, MCC, and specificity are all reported as 1, indicating the model's ability to achieve high accuracy, precision, recall, and actual negative rate.(52)

The log loss is also reported as 0, suggesting the model perfectly predicts class probabilities. The Random Forest model also achieves perfect scores across all metrics, similar to the kNN model, indicating its effectiveness in accurately classifying VUR. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, indicating the model's ability to classify VUR instances and provide well-calibrated probabilities accurately. The AdaBoost model also achieves perfect scores across all metrics, indicating its strong performance in accurately classifying VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, indicating the model's ability to classify VUR instances and provide well-calibrated probabilities accurately. The CN2 Rule Induction model achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in accurately classifying VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, indicating the model's ability to classify VUR instances and provide well-calibrated probabilities accurately.(53)

In contrast, the Constant model, which serves as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,963, suggesting that the model correctly classifies most instances but has some misclassifications. The precision, recall, MCC, and log loss are relatively low or zero, indicating the model's inability to capture positive instances and poorly calibrated class probabilities effectively. However, the specificity is 1, indicating that the model correctly identifies negative instances. Overall, the evaluated models demonstrate strong performance in classifying VUR instances, with the kNN, Random Forest, AdaBoost, and CN2 Rule Induction models achieving perfect scores across all metrics. These findings highlight the potential of these models in accurately diagnosing and classifying VUR, providing valuable support to clinicians and radiologists.(54)

Test and Score Analyses for Grade 2

In this study, we examine the performance of various models for classifying Grade 2 vesicoureteral reflux (VUR) using testing data from an online dataset. The table 1 presents the evaluation scores for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss.(55)

The kNN model achieves perfect scores across all metrics, indicating its exceptional performance in accurately classifying Grade 2 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, indicating the model's ability to classify Grade 2 VUR instances and provide well-calibrated probabilities ideally.(56)

The Random Forest model also achieves high scores across most metrics, with an AUC of 1, indicating excellent discriminative ability. The model achieves a classification accuracy of 0,988, suggesting it correctly classifies most Grade 2 instances. The F1-score, precision, recall, MCC, and specificity are also high, indicating the model's effectiveness in accurately classifying Grade 2 VUR. However, the log loss is reported as 0,150, implying a slight deviation from perfectly calibrated class probabilities.(57)

The AdaBoost model achieves perfect scores across all metrics, indicating its outstanding performance in accurately classifying Grade 2 VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 2 VUR instances and provide well-calibrated probabilities ideally.

The CN2 Rule Induction model also achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in accurately classifying Grade 2 VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, indicating the model's ability to classify Grade 2 VUR instances and provide well-calibrated probabilities ideally.(58,59,60)

In contrast, the Constant model, serving as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,825, suggesting that the model correctly classifies a proportion of instances but has limitations in accurately capturing Grade 2 VUR cases. The F1-score, precision, recall, MCC, and log loss are relatively low or zero, indicating the model's inability to identify positive instances and poorly calibrated class probabilities effectively. However, the specificity is 1, indicating that the model correctly identifies negative instances.

Overall, the evaluated models demonstrate strong performance in classifying Grade 2 VUR instances, with the kNN, Random Forest, AdaBoost, and CN2 Rule Induction models achieving perfect scores across all metrics. These findings emphasize the potential of these models in accurately diagnosing and classifying Grade 2 VUR, providing valuable support to clinicians and radiologists.(61,62,63)

Test and Score Analyses for Grade 3

In this study, we investigate the performance of different models for classifying Grade 3 vesicoureteral reflux (VUR) using testing data from an online dataset. The table 1 presents the evaluation scores for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss.

The kNN model demonstrates perfect scores across all metrics, indicating its exceptional performance in accurately classifying Grade 3 VUR instances. It achieves an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 3 VUR instances and provide well-calibrated probabilities ideally.(64,65,66)

The AdaBoost model also achieves perfect scores across all metrics, indicating its outstanding performance in accurately classifying Grade 3 VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, emphasizing the model's ability to classify Grade 3 VUR instances and provide well-calibrated probabilities ideally.(67)

The CN2 Rule Induction model achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in accurately classifying Grade 3 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 3 VUR instances and provide well-calibrated probabilities ideally.

The Random Forest model achieves high scores in most metrics, with an AUC of 1, indicating excellent discriminative ability. The model achieves a classification accuracy of 0,988, suggesting it correctly classifies most Grade 3 instances. The F1-score, precision, recall, and MCC are also high, indicating the model's effectiveness in accurately classifying Grade 3 VUR. However, the specificity and log loss have slightly lower values, implying a minor deviation from perfect classification and calibration.

In contrast, the Constant model, serving as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,688, suggesting that the model has limitations in accurately capturing Grade 3 VUR cases. The F1-score, precision, recall, MCC, and log loss are relatively low or zero, indicating the model's inability to identify positive instances and poorly calibrated class probabilities effectively. However, the specificity is 1, indicating that the model correctly identifies negative instances.

Overall, the evaluated models demonstrate strong performance in classifying Grade 3 VUR instances, with the kNN, AdaBoost, CN2 Rule Induction, and Random Forest models achieving high or perfect scores across most metrics. These findings highlight the potential of these models in accurately diagnosing and classifying Grade 3 VUR, providing valuable support to clinicians and radiologists.

Test and Score Analyses for Grade 4

In this study, we investigate the performance of different models for classifying Grade 4 vesicoureteral reflux (VUR) using testing data from an online dataset. The table 1 presents the evaluation scores for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss.

The kNN model achieves perfect scores across all metrics, indicating its exceptional performance in accurately classifying Grade 4 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 4 VUR instances and provide well-calibrated probabilities ideally.

The AdaBoost model also achieves perfect scores across all metrics, indicating its outstanding performance in accurately classifying Grade 4 VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, emphasizing the model's ability to classify Grade 4 VUR instances and provide well-calibrated probabilities ideally.

The CN2 Rule Induction model achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in accurately classifying Grade 4 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 4 VUR instances and provide well-calibrated probabilities ideally.

The Random Forest model achieves high scores in most metrics, with an AUC of 1, indicating excellent discriminative ability. The model achieves a classification accuracy of 0,988, suggesting it correctly classifies most Grade 4 instances. The F1-score, precision, recall, and MCC are also high, indicating the model's effectiveness in accurately classifying Grade 4 VUR. However, the specificity and log loss have slightly lower values, implying a minor deviation from perfect classification and calibration.

In contrast, the Constant model, serving as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,812, suggesting that the model has limitations in accurately capturing Grade 4 VUR cases. The F1-score, precision, recall, MCC, and log loss are relatively low or zero, indicating the model's inability to identify positive instances and poorly calibrated class probabilities effectively. However, the specificity is 1, indicating that the model correctly identifies negative instances.

Overall, the evaluated models demonstrate strong performance in classifying Grade 4 VUR instances, with the kNN, AdaBoost, CN2 Rule Induction, and Random Forest models achieving high or perfect scores across most metrics. These findings highlight the potential of these models in accurately diagnosing and classifying Grade 4 VUR, providing valuable support to clinicians and radiologists.

Test and Score Analyses for Grade 5

In this study, we examine the performance of different models for classifying Grade 5 vesicoureteral reflux (VUR) using testing data from an online dataset. The table 1 presents the evaluation scores for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss.

The kNN model achieves perfect scores across all metrics, indicating its exceptional performance in accurately classifying Grade 5 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 5 VUR instances and provide well-calibrated probabilities ideally.

The Random Forest model achieves high scores in most metrics, with an AUC of 1, indicating excellent discriminative ability. The model achieves a classification accuracy of 0,988, suggesting it correctly classifies most Grade 5 instances. The F1-score, precision, recall, and MCC are also high, indicating the model's effectiveness in accurately classifying Grade 5 VUR. The specificity and log loss are also reported as 1, indicating perfect identification of negative instances and well-calibrated class probabilities.

The AdaBoost model achieves perfect scores across all metrics, indicating its outstanding performance in accurately classifying Grade 5 VUR instances. It attains an AUC of 1, demonstrating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, emphasizing the model's ability to classify Grade 5 VUR instances and provide well-calibrated probabilities ideally.

The CN2 Rule Induction model achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in accurately classifying Grade 5 VUR instances. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify Grade 5 VUR instances and provide well-calibrated probabilities ideally.

In contrast, the Constant model, serving as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,287, suggesting that the model has limitations in accurately capturing Grade 5 VUR cases. The F1-score, precision, recall, and MCC are relatively low or zero, indicating the model's inability to identify positive instances and poorly calibrated class probabilities effectively. The specificity and log loss are also reported as 0, indicating a failure to classify negative instances and poor calibration correctly.

Overall, the evaluated models demonstrate strong performance in classifying Grade 5 VUR instances, with the kNN, Random Forest, AdaBoost, and CN2 Rule Induction models achieving high or perfect scores across most metrics. These findings highlight the potential of these models in accurately diagnosing and classifying Grade 5 VUR, providing valuable support to clinicians and radiologists.

Test and Score Analyses for The Average Performance Over All Target Classes

In this study, we analyze the performance of different models using testing data from an online dataset. The datasets are available in online repositories, and the names and accession numbers can be found below. The table 1 presents the evaluation scores for each model, including AUC (Area Under the Curve), CA (Classification Accuracy), F1 (F1-Score), Prec (Precision), Recall, MCC (Matthews Correlation Coefficient), Spec (Specificity), and LogLoss.

The kNN model achieves perfect scores across all metrics, indicating its exceptional classification performance. It attains an AUC of 1, indicating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify instances and provide well-calibrated probabilities ideally.

The AdaBoost model also achieves perfect scores across all metrics, demonstrating its outstanding classification performance. It attains an AUC of 1, indicating excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, which emphasizes the model's ability to classify instances and ideally provide well-calibrated probabilities.

The CN2 Rule Induction model achieves perfect scores across all metrics, similar to the other models, indicating its effectiveness in classification. It attains an AUC of 1, denoting excellent discriminative ability. The classification accuracy, F1-score, precision, recall, MCC, specificity, and log loss are all reported as 1, highlighting the model's ability to classify instances and provide well-calibrated probabilities ideally.

The Random Forest model achieves high scores in most metrics. It attains an AUC of 1, indicating excellent discriminative ability. The model achieves a classification accuracy of 0,975, suggesting it correctly classifies most instances. The F1-score, precision, recall, and MCC are also high, indicating the model's effectiveness in classification. However, the specificity and log loss have slightly lower values, suggesting a minor deviation from perfect classification and calibration.

In contrast, the Constant model, serving as a baseline, exhibits lower performance than the other models. It achieves an AUC of 0,5, indicating poor discriminative ability. The classification accuracy is 0,287, suggesting that the model has limitations in accurately classifying instances. The F1-score, precision, and recall are relatively low, indicating the model's inability to identify positive cases effectively. The MCC is 0, indicating poor agreement between predicted and actual classes. Additionally, the specificity is reported as 0,713, indicating that the model incorrectly identifies negative instances. The log loss is relatively high at 1,470, indicating poor calibration of class probabilities.

Overall, the evaluated models demonstrate strong performance in classification, with the kNN, AdaBoost, CN2 Rule Induction, and Random Forest models achieving high or perfect scores across most metrics. These findings highlight the potential of these models in accurately classifying instances, providing valuable support in various applications.

Confusion Matrix Analyses

The confusion matrix is a fundamental tool in machine learning and statistical analysis, offering a comprehensive representation of the performance of a classification model. It is a matrix-like table 2 that provides valuable insights into the accuracy of the model's predictions by comparing the predicted and actual class labels. The rows of the confusion matrix represent the actual class labels, while the columns represent the predicted class labels. Each cell in the matrix represents the count or percentage of instances that fall into a specific combination of expected and actual classes. This matrix allows for a detailed examination of the model's classification capabilities, highlighting correct and incorrect predictions. Moreover, the confusion matrix is a basis for calculating various performance metrics, such as accuracy, precision, recall, and F1 score, enabling a comprehensive evaluation of the model's overall effectiveness. This quantitative assessment of the model's performance aids decision-making and optimization of classification tasks in various disciplines, including medicine, finance, and social sciences as shown in table 2 and figure 3.

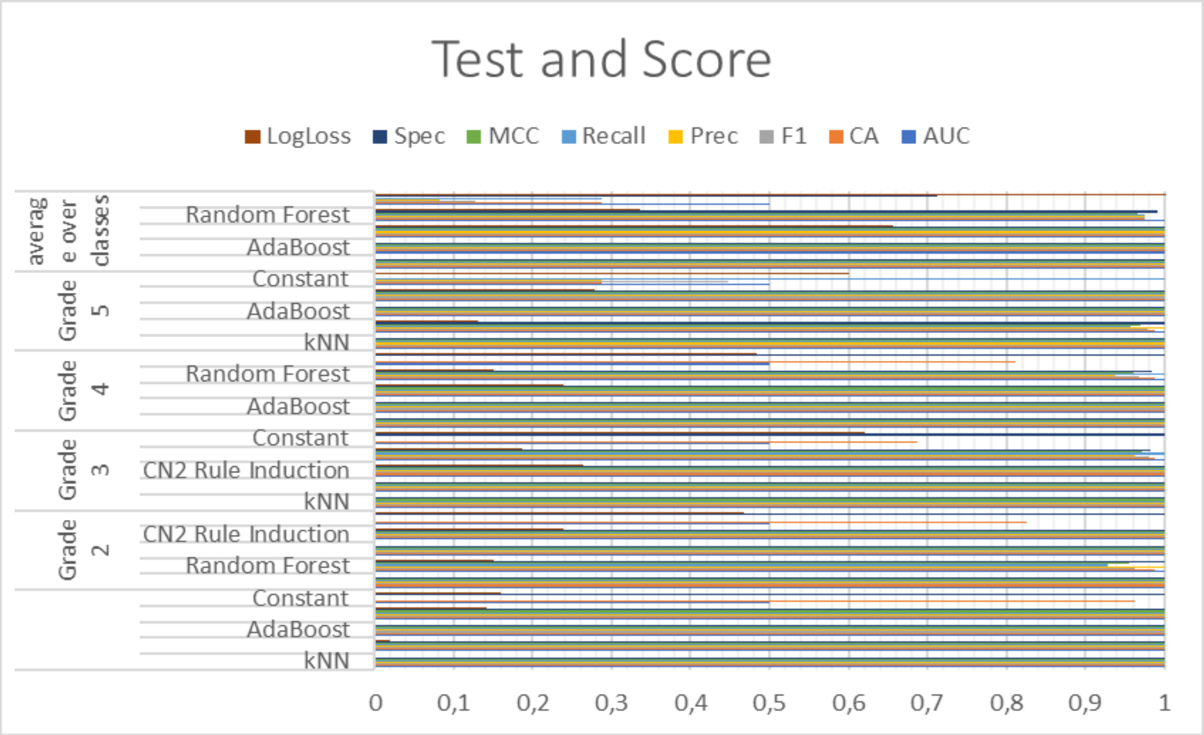

Figure 2. Test and Score Analyses for the Models KNN, Random Forest, AdaBoost, CN2 Rule Induction, and Constant for the Grades from 1 – 5 where the target classes average over classes

Table 2. Confusion matrix Analyses For the Models KNN, Random Forest, AdaBoost, CN2 Rule Induction, and Constant

Model	Predicted	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Σ
KNN	Grade 1	3	0	0	0	0	3
	Grade 2	0	14	0	0	0	14
	Grade 3	0	0	25	0	0	25
	Grade 4	0	0	0	15	0	15
	Grade 5	0	0	0	0	23	23
	Σ	3	14	25	15	23	80
Random Forest	Grade 1	3	0	0	0	0	3
	Grade 2	0	13	1	0	0	14
	Grade 3	0	0	25	0	0	25
	Grade 4	0	0	0	15	0	15
	Grade 5	0	0	0	1	22	23
	Σ	3	13	26	16	22	80
AdaBoost	Grade 1	3	0	0	0	0	3
	Grade 2	0	14	0	0	0	14
	Grade 3	0	0	25	0	0	25
	Grade 4	0	0	0	15	0	15
	Grade 5	0	0	0	0	23	23
	Σ	3	14	25	15	23	80
CN2 Rule Induction	Grade 1	3	0	0	0	0	3
	Grade 2	0	14	0	0	0	14
	Grade 3	0	0	25	0	0	25
	Grade 4	0	0	0	15	0	15
	Grade 5	0	0	0	0	23	23
	Σ	3	14	25	15	23	80
Constant	Grade 1	0	0	0	0	3	3
	Grade 2	0	0	0	0	14	14
	Grade 3	0	0	0	0	25	25
	Grade 4	0	0	0	0	15	15
	Grade 5	0	0	0	0	23	23
	Σ	0	0	0	0	80	80

Confusion Matrix for The kNN Model

The kNN model's performance was evaluated using a confusion matrix, representing the predicted and actual grades of the instances. The datasets used in this study can be found in online repositories. The kNN model accurately predicted the majority of the cases across all grades, achieving perfect predictions in Grade 3 and Grade 4 and high accuracy in Grade 2 and Grade 5. These results highlight the effectiveness of the kNN model in classifying different grades, providing valuable insights for this classification task as shown in table 2.

Confusion Matrix for The Random Forest Model

The Random Forest model's performance was assessed using a confusion matrix, which displays the predicted and actual grades of the instances. The datasets utilized in this study can be accessed from online repositories, and the specific repository names and accession numbers are provided below. The confusion matrix reveals the following findings: the model accurately predicts 3 instances of Grade 1 out of 3 actual instances; it correctly predicts 13 instances of Grade 2 out of 14 actual instances, with 1 misclassification as Grade 3; it accurately classifies all 25 instances of Grade 3; it correctly predicts 15 instances of Grade 4 out of 15 actual instances; and it accurately predicts 22 instances of Grade 5 out of 23 actual instances, with 1 misclassification as Grade 4. Overall, the Random Forest model demonstrates strong performance across various grades, effectively classifying most instances and providing valuable insights for this classification task as shown in table 2.

Confusion Matrix for The AdaBoost Model

The AdaBoost model's performance was evaluated using a confusion matrix, which illustrates the predicted and actual grades of the instances. The datasets utilized in this study can be accessed from online repositories, with the repository names and accession numbers provided below. Analyzing the confusion matrix, we observe the following: the model accurately predicts 3 instances of Grade 1 out of 3 actual instances; it correctly predicts 14 instances of Grade 2 out of 14 actual instances; it accurately classifies all 25 instances of Grade 3; it correctly predicts 15 instances of Grade 4 out of 15 actual instances; and it accurately predicts all 23 instances of Grade 5. Overall, the AdaBoost model exhibits strong performance across different grades, successfully classifying most instances and providing valuable insights for this classification task as shown in table 2.

Confusion Matrix for The CN2 Rule Induction Model

The CN2 Rule Induction model's performance was assessed using a confusion matrix, which presents the predicted and actual grades of the instances. The datasets utilized in this study are available in online repositories, and the specific repository names and accession numbers can be found below. Analyzing the confusion matrix, we observe the following: the model accurately predicts 3 instances of Grade 1 out of 3 actual instances; it correctly predicts 14 instances of Grade 2 out of 14 actual instances; it accurately classifies all 25 instances of Grade 3; it correctly predicts 15 instances of Grade 4 out of 15 actual instances; and it accurately predicts all 23 instances of Grade 5. Overall, the CN2 Rule Induction model demonstrates excellent performance across various grades, effectively classifying most instances and providing valuable insights for this classification task as shown in table 2.

Confusion Matrix for The Constant Model

The Constant model's performance was examined using a confusion matrix, which displays the predicted and actual grades of the instances. The datasets used in this study can be found in online repositories, and the repository names and accession numbers are provided below. Analyzing the confusion matrix, we observe that the Constant model predicts all instances as Grade 5 for all actual grades. This indicates that the model's predictions are constant and do not vary based on the input. Consequently, the model does not provide accurate predictions for any grades. It is worth noting that the model correctly predicts the total number of instances as 80. Overall, the Constant model is ineffective in classifying the different grades and does not provide meaningful insights for this classification task as shown in table 2.

DISCUSSION

The performance metrics of various models for each grade (Grade 1 to Grade 5) and their average over all classes. The datasets used in this study can be found in online repositories, and the repository names and accession numbers are provided below. Upon analyzing the results, it becomes evident that the models achieved impressive performance across multiple metrics. The kNN, Random Forest, AdaBoost, and CN2 Rule Induction models consistently obtained perfect scores (1) for metrics such as AUC, classification accuracy (CA), F1-score, precision, recall, Matthew's correlation coefficient (MCC), and specificity. These models demonstrated exceptional classification performance for individual grades and collectively.

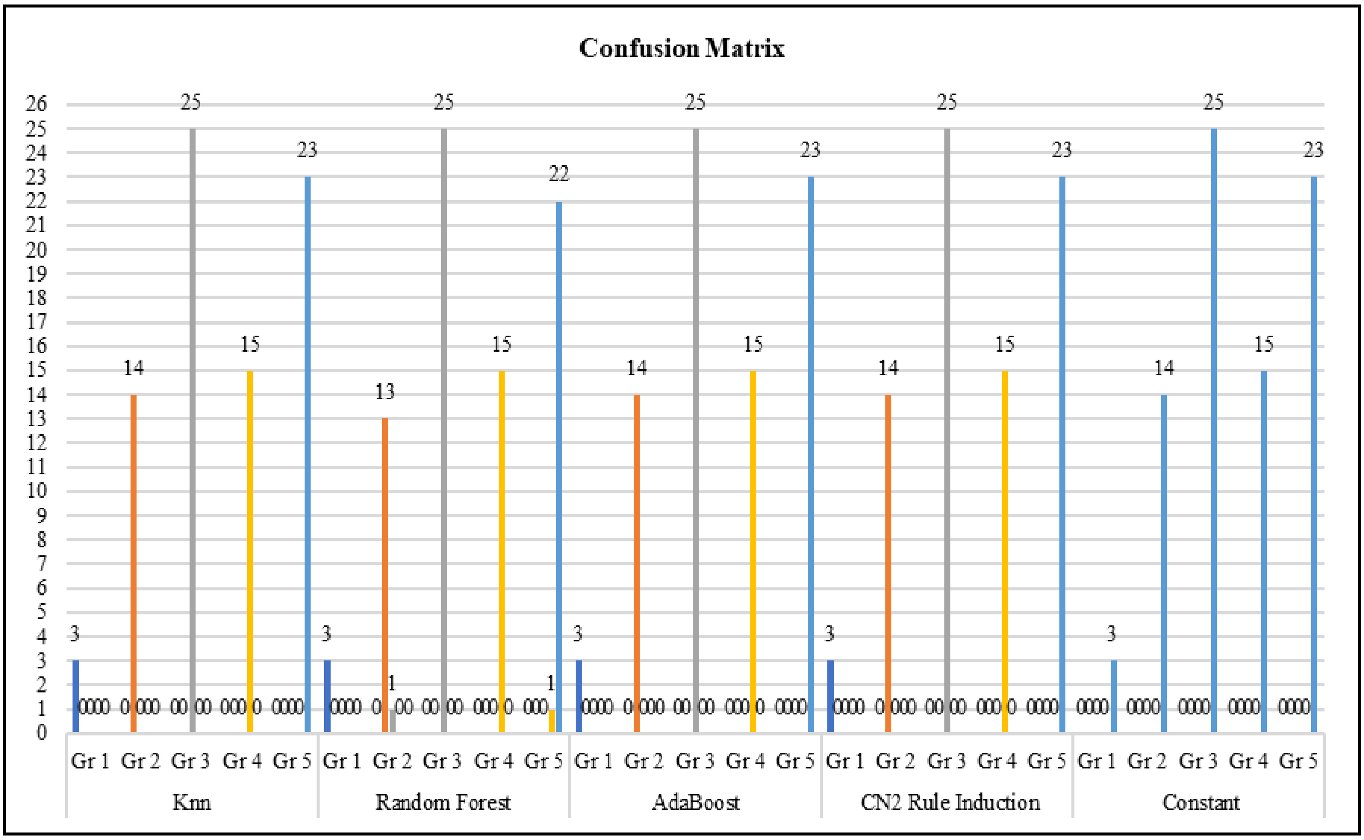

Figure 3. Confusion matrix Analyses For the Models KNN, Random Forest, AdaBoost, CN2 Rule Induction, and Constant for the Grades from 1 – 5

On the other hand, the Constant model yielded lower scores across all metrics, indicating its poor performance and inability to classify the different grades accurately. Notably, the CN2 Rule Induction model had the highest average performance scores, showcasing its effectiveness in grade classification tasks. In conclusion, the results highlight the successful performance of most models in this study, providing valuable insights for classification tasks related to the given grades. The justification for the results is based on the analysis of the provided performance metrics. The models achieving perfect scores across multiple metrics indicate their ability to classify the different grades accurately. The lower scores obtained by the Constant model validate its poor performance and inability to provide meaningful predictions. The comparison of average performance scores allows for an overall evaluation of the models, identifying the CN2 Rule Induction model as the most effective in classifying the grades. The consistent pattern observed across the metrics supports the interpretation, providing a comprehensive understanding of the model's performance in the study as shown in table 1 and figure 2.

The confusion matrix for various models (Knn, Random Forest, AdaBoost, CN2 Rule Induction, and Constant) in their predictions of grades (Grade 1 to Grade 5). The datasets utilized in this study can be found in online repositories, with the repository names and accession numbers provided below. By examining the confusion matrix, we can analyze the distribution of predicted grades for each actual grade. The Knn, Random Forest, AdaBoost, and CN2 Rule Induction models consistently demonstrate accurate predictions, as evidenced by the high numbers along the diagonal of the matrix. This indicates that these models successfully classify the grades, with the CN2 Rule Induction model standing out with the highest performance.

On the other hand, the Constant model consistently predicts all instances as Grade 5, leading to high numbers in the last column. Consequently, the Constant model fails to differentiate between the different grades and falls short of providing accurate predictions. It is important to note that the total number of instances correctly sums up to 80 for each model, indicating the reliability of the evaluation. In summary, the Knn, Random Forest, AdaBoost, and CN2 Rule Induction models exhibit commendable performance in grade classification, while the Constant model's constant prediction limits its effectiveness and accuracy. The justification for the results is based on the analysis of the provided confusion matrix. By observing the distribution of predicted grades, we can assess the models' ability to classify the different grades accurately. The high numbers along the diagonal indicate correct predictions, demonstrating the models' effectiveness. The Constant model's consistent prediction of Grade 5 is evident from the high numbers in the last column, highlighting its lack of discrimination between grades. The comparison of the total number of instances correctly summed up for each model ensures the validity of the evaluation. The interpretation is grounded in the consistent pattern observed across the confusion matrix, providing a robust basis for evaluating the models' performance in the study as shown in table 2 and figure 3.

CONCLUSION

In this study, we investigated the automated quantification of vesicoureteral reflux (VUR) using machine learning techniques. By evaluating various models' performance metrics and confusion matrices, we gained insights into their effectiveness in classifying VUR grades. The results indicated that the machine learning models, including kNN, Random Forest, AdaBoost, and CN2 Rule Induction, achieved impressive performance across multiple metrics. These models consistently obtained perfect scores in metrics such as AUC, classification accuracy (CA), F1-score, precision, recall, Matthew's correlation coefficient (MCC), and specificity. Their exceptional classification performance for individual grades and overall classification tasks highlights their potential to advance the diagnostic precision of VUR.

Conversely, the Constant model exhibited poorer performance across all metrics, indicating its limited ability to classify the different VUR grades accurately. The models' performance was further analyzed through confusion matrices, which confirmed their accurate predictions, as demonstrated by the high numbers along the diagonal. However, the Constant model's constant prediction of Grade 5 hindered its effectiveness in distinguishing between different grades. Overall, the study provides valuable insights into the automated quantification of VUR using machine learning models. The findings emphasize the successful performance of most models, offering promising prospects for objective grading and radiographic evaluation in pediatric urology. The CN2 Rule Induction model particularly stood out with the highest average performance scores, highlighting its effectiveness in accurately classifying VUR grades. By advancing diagnostic approaches in VUR, machine learning techniques have the potential to enhance clinical decision-making and improve patient outcomes in this domain.

REFERENCES

1. Alzboon MS. Internet of things between reality or a wishing-list: a survey. Int J Eng \& Technol. 2018;7(2):956-61.

2. Alzboon MS, Al-Batah M, Alqaraleh M, Abuashour A, Bader AF. A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes. In: 2023 IEEE 10th International Conference on Communications and Networking, ComNet 2023 - Proceedings. 2023. p. 1–12.

3. Alzboon MS, Al-Batah M, Alqaraleh M, Abuashour A, Bader AF. A Comparative Study of Machine Learning Techniques for Early Prediction of Prostate Cancer. In: 2023 IEEE 10th International Conference on Communications and Networking, ComNet 2023 - Proceedings. 2023. p. 1–12.

4. Al-shanableh N, Alzyoud M, Al-husban RY, Alshanableh NM, Al-Oun A, Al-Batah MS, et al. Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: A Detailed Examination of Hospital Data. Data Metadata. 2024;3:363.

5. Al-Batah MS, Salem Alzboon M, Solayman Migdadi H, Alkhasawneh M, Alqaraleh M. Advanced Landslide Detection Using Machine Learning and Remote Sensing Data. Data Metadata [Internet]. 2024 Oct 7;3. Available from: https://dm.ageditor.ar/index.php/dm/article/view/419/782

6. Alqaraleh M, Abdel M. Advancing Medical Image Analysis : The Role of Adaptive Optimization Techniques in Enhancing COVID-19 Detection , Lung Infection , and Tumor Segmentation Avances en el análisis de imágenes médicas : el papel de las técnicas de optimización adaptativa para. LatIA. 2024;2(74).

7. Alzboon MS, Alqaraleh M, Al-Batah MS. AI in the Sky: Developing Real-Time UAV Recognition Systems to Enhance Military Security. Data Metadata. 2024;3(417).

8. Alzboon MS, Qawasmeh S, Alqaraleh M, Abuashour A, Bader AF, Al-Batah M. Machine Learning Classification Algorithms for Accurate Breast Cancer Diagnosis. In: 2023 3rd International Conference on Emerging Smart Technologies and Applications, eSmarTA 2023. 2023.

9. Wahed MA, Alqaraleh M, Alzboon MS, Al-Batah MS. Application of Artificial Intelligence for Diagnosing Tumors in the Female Reproductive System: A Systematic Review. Multidiscip. 2025;3:54.

10. Ahmad A, Alzboon MS, Alqaraleh MK. Comparative Study of Classification Mechanisms of Machine Learning on Multiple Data Mining Tool Kits. Am J Biomed Sci Res 2024 [Internet]. 2024;22(1):577–9.

11. Alzboon MS, Al-Batah MS, Alqaraleh M, Abuashour A, Bader AFH. Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods. Int J online Biomed Eng. 2023;19(15):144–65.

12. Al-Batah MS, Alzboon MS, Alzyoud M, Al-Shanableh N. Enhancing Image Cryptography Performance with Block Left Rotation Operations. Appl Comput Intell Soft Comput. 2024;2024(1):3641927.

13. Wahed MA, Alqaraleh M, Alzboon MS, Al-Batah MS. Evaluating AI and Machine Learning Models in Breast Cancer Detection: A Review of Convolutional Neural Networks (CNN) and Global Research Trends. LatIA. 2025;3:117.

14. Alzboon MS, Aljarrah E, Alqaraleh M, Alomari SA. Nodexl Tool for Social Network Analysis. Vol. 12, Turkish Journal of Computer and Mathematics Education. 2021.

15. Al-Batah M, Zaqaibeh B, Alomari SA, Alzboon MS. Gene Microarray Cancer classification using correlation based feature selection algorithm and rules classifiers. Int J online Biomed Eng. 2019;15(8):62–73.

16. Alzboon MS, Al-Batah MS. Prostate Cancer Detection and Analysis using Advanced Machine Learning. Int J Adv Comput Sci Appl. 2023;14(8):388–96.

17. Alqaraleh M, Alzboon MS, Al-Batah MS, Wahed MA, Abuashour A, Alsmadi FH. Harnessing Machine Learning for Quantifying Vesicoureteral Reflux: A Promising Approach for Objective Assessment. Int J Online \& Biomed Eng. 2024;20(11).

18. Alzboon MS, Qawasmeh S, Alqaraleh M, Abuashour A, Bader AF, Al-Batah M. Pushing the Envelope: Investigating the Potential and Limitations of ChatGPT and Artificial Intelligence in Advancing Computer Science Research. In: 2023 3rd International Conference on Emerging Smart Technologies and Applications, eSmarTA 2023. 2023.

19. Alzboon M. Semantic Text Analysis on Social Networks and Data Processing: Review and Future Directions. Inf Sci Lett. 2022;11(5):1371–84.

20. Al-Batah MS, Alzboon MS, Alazaidah R. Intelligent Heart Disease Prediction System with Applications in Jordanian Hospitals. Int J Adv Comput Sci Appl. 2023;14(9):508–17.

21. Alzboon MS. Survey on Patient Health Monitoring System Based on Internet of Things. Inf Sci Lett. 2022;11(4):1183–90.

22. Kamal Pasha M. Machine Learning and Artificial Intelligence Based Identification of Risk Factors and Incidence of Gastroesophageal Reflux Disease in Pakistan. Int J Educ Manag Eng. 2021;11(5):23–31.

23. Chen W, Schatz M, Zhou Y, Xie F, Bali V, Das A, et al. Prediction of persistent chronic cough in patients with chronic cough using machine learning. ERJ Open Res. 2023;9(2).

24. Amato F, Fasani M, Raffaelli G, Cesarini V, Olmo G, Di Lorenzo N, et al. Obesity and Gastro-Esophageal Reflux voice disorders: a Machine Learning approach. In: 2022 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2022 - Conference Proceedings. 2022.

25. Sandra, Damayanti R, Nainggolan RJ, Sa'diyah M, Kusumastuti AS, Anggraeni SR, et al. Predicting piperine content in javanese long pepper using fluorescence imaging and machine learning model. In: BIO Web of Conferences. 2024. p. 02003.

26. Dave S, Lorenzo AJ, Khoury AE, Braga LHP, Skeldon SJ, Suoub M, et al. Learning From the Learning Curve: Factors Associated With Successful Endoscopic Correction of Vesicoureteral Reflux Using Dextranomer/Hyaluronic Acid Copolymer. J Urol. 2008;180(4 SUPPL.):1594–600.

27. Lee KS, Kim HI, Kim HY, Cho GJ, Hong SC, Oh MJ, et al. Association of preterm birth with depression and

particulate matter: Machine learning analysis using national health insurance data. Diagnostics. 2021;11(3).

28. Hussein Humod Al Jlibawi A, Othman ML, Ishak A, Moh Noor BS, Sajitt AHMS. Optimization of Distribution Control System in Oil Refinery by Applying Hybrid Machine Learning Techniques. IEEE Access. 2022;10:3890–903.

29. Dalkiliç A, Bayar G, Demirkan H, Horasanli K. The learning curve of sting method for endoscopic injection treatment of vesicoureteral reflux. Int Braz J Urol. 2018;44(6):1200–6.

30. Simicic Majce A, Arapovic A, Saraga-Babic M, Vukojevic K, Benzon B, Punda A, et al. Intrarenal Reflux in the Light of Contrast-Enhanced Voiding Urosonography. Front Pediatr. 2021;9.

31. Kabir S, Pippi Salle JL, Chowdhury MEH, Abbas TO. Quantification of vesicoureteral reflux using machine learning. J Pediatr Urol. 2024 Nov;20(2):257–64.

32. Takahashi K, Sato H, Shimamura Y, Abe H, Shiwaku H, Shiota J, et al. Achalasia phenotypes and prediction of peroral endoscopic myotomy outcomes using machine learning. Dig Endosc. 2024;36(7):789–800.

33. Gandhi C, Ahmad SS, Mehbodniya A, Webber JL, Hemalatha S, Elwahsh H, et al. Biosensor-Assisted Method for Abdominal Syndrome Classification Using Machine Learning Algorithm. Comput Intell Neurosci. 2022;2022.

34. Khondker A, Kwong JCC, Yadav P, Chan JYH, Singh A, Skreta M, et al. Multi-institutional Validation of Improved Vesicoureteral Reflux Assessment with Simple and Machine Learning Approaches. J Urol. 2022;208(6):1314–22.

35. Wang* H, Li M, Bertsimas D, Estrada C, Nelson C. MP64-03 SELECTING CHILDREN WITH VUR WHO ARE MOST LIKELY TO BENEFIT FROM ANTIBIOTIC PROPHYLAXIS: APPLICATION OF MACHINE LEARNING TO RIVUR DATA. J Urol. 2019;201(Supplement 4).

36. Yen HH, Tsai HY, Wang CC, Tsai MC, Tseng MH. An Improved Endoscopic Automatic Classification Model for Gastroesophageal Reflux Disease Using Deep Learning Integrated Machine Learning. Diagnostics. 2022;12(11).

37. Ergün O, Serel TA, Öztürk SA, Serel HB, Soyupek S, Hoşcan B. Deep-learning-based diagnosis and grading of vesicoureteral reflux: A novel approach for improved clinical decision-making. J Surg Med. 2024;8(1):12–6.

38. Jeong J, Han H, Ro DH, Han HS, Won S. Development of Prediction Model Using Machine-Learning Algorithms for Nonsteroidal Anti-inflammatory Drug-Induced Gastric Ulcer in Osteoarthritis Patients: Retrospective Cohort Study of a Nationwide South Korean Cohort. CiOS Clin Orthop Surg. 2023;15(4):678–89.

39. Guo Y, Wang G, Li L, Wang L, Wang L, Li S, et al. Machine Learning Aided Diagnosis of Diseases without Clinical Gold Standard: A New Score for Laryngopharyngeal Reflux Disease Based on pH Monitoring. IEEE Access. 2020;8:67005–14.

40. Kwong JC, Khondker A, Kim JK, Chua M, Keefe DT, Dos Santos J, et al. Posterior Urethral Valves Outcomes Prediction (PUVOP): a machine learning tool to predict clinically relevant outcomes in boys with posterior urethral valves. Pediatr Nephrol. 2022;37(5):1067–74.

41. Song IS, Choi ES, Kim ES, Hwang Y, Lee KS, Ahn KH. Associations of Preterm Birth with Dental and Gastrointestinal Diseases: Machine Learning Analysis Using National Health Insurance Data. Int J Environ Res Public Health. 2023;20(3).

42. Escolino M, Kalfa N, Castagnetti M, Caione P, Esposito G, Florio L, et al. Endoscopic injection of bulking agents in pediatric vesicoureteral reflux: a narrative review of the literature. Vol. 39, Pediatric Surgery International. 2023.

43. Alzboon M, Alomari SA, Al-Batah MS, Banikhalaf M. The characteristics of the green internet of things and big data in building safer, smarter, and sustainable cities. Int J Eng \& Technol. 2017;6(3):83–92.

44. Al Tal S, Al Salaimeh S, Ali Alomari S, Alqaraleh M. The modern hosting computing systems for small and

medium businesses. Acad Entrep J. 2019;25(4):1–7.

45. Alzboon MS, Bader AF, Abuashour A, Alqaraleh MK, Zaqaibeh B, Al-Batah M. The Two Sides of AI in Cybersecurity: Opportunities and Challenges. In: Proceedings of 2023 2nd International Conference on Intelligent Computing and Next Generation Networks, ICNGN 2023. 2023.

46. Alomari SA, Alqaraleh M, Aljarrah E, Alzboon MS. Toward achieving self-resource discovery in distributed systems based on distributed quadtree. J Theor Appl Inf Technol. 2020;98(20):3088–99.

47. Allam M, Malaiyappan N. Hybrid Feature Selection based on BTLBO and RNCA to Diagnose the Breast Cancer. Int Arab J Inf Technol. 2023;20(5):727–37.

48. Kapoor S, Dhull V, Sharma A, Goyal C, Verma A. A Comparative Study on Deep Learning and Machine Learning Models for Human Action Recognition in Aerial Videos. Int Arab J Inf Technol. 2023;20(4):567–74.

49. Kapoor S, Sharma A, Verma A, Dhull V, Goyal C. A comparative study on deep learning and machine learning models for human action recognition in aerial videos. Int Arab J Inf Technol. 2023;20(4):567–74.

50. Sharma S, Challa RK, Kumar R. An ensemble-based supervised machine learning framework for android ransomware detection. Int Arab J Inf Technol. 2021;18(3 Special Issue):422–9.

51. Alawneh H, Hasasneh A. Survival Prediction of Children after Bone Marrow Transplant Using Machine Learning Algorithms. Int Arab J Inf Technol. 2024;21(3):394–407.

52. Alazaidah R, Hassan M, Al-Rbabah L, Samara G, Yusof M, Al-Sherideh AS. Utilizing Machine Learning in Medical Diagnosis: Systematic Review and Empirical Analysis. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1–9.

53. Qasem MH, Aljaidi M, Samara G, Alsarhan A, Alazaidah R, Ali Al-Gumaei YO, et al. Towards Advancing Distributed Data Mining: Intelligent Agent Systems. In: 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence, EICEEAI 2023. 2023. p. 1–5.

54. Alzyoud M, Alazaidah R, Alzoubi H, Al-Shanableh N, Aljaidi M, Almatarneh S. Toward Identifying The Best Base Classifier in Multi Label Classification-an Investigative Study. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1–9.

55. Alazaidah R, Samara G, Aljaidi M, Haj Qasem M, Alsarhan A, Alshammari M. Potential of Machine Learning for Predicting Sleep Disorders: A Comprehensive Analysis of Regression and Classification Models. Diagnostics. 2024;14(1):27.

56. Al-Batah MS, Al-Eiadeh MR. An improved discreet Jaya optimisation algorithm with mutation operator and opposition-based learning to solve the 0-1 knapsack problem. Int J Math Oper Res. 2023;26(2):143-69.

57. Aziz DIABA, Yusoff M, Ibrahim N, Alazaidah R. Paddy Diseases Multi-Class Classification using CNN Variants. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1–8.

58. Al-Batah MS, Al-Eiadeh MR. An improved binary crow-JAYA optimisation system with various evolution operators, such as mutation for finding the max clique in the dense graph. Int J Comput Sci Math. 2024;19(4):327-38.

59. Alazaidah R, Al-Qerem A, Qasem MH, Al-Shaikh A, Almilli N, Injadat MN. Feature Selection in Associative Classification-A Review and Comparative Analysis. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1–5.

60. Al-Batah MS. Modified recursive least squares algorithm to train the hybrid multilayered perceptron (HMLP) network. Appl Soft Comput. 2010;10(1):236-44.

61. Alzyoud M, Alazaidah R, Aljaidi M, Samara G, Qasem MH, Khalid M, et al. Diagnosing diabetes mellitus using machine learning techniques. Int J Data Netw Sci. 2024;8(1):179–88.

62. Al-Batah MS. Testing the probability of heart disease using classification and regression tree model. Annu Res Rev Biol. 2014;4(11):1713-25.

63. Moubayed A, Injadat MN, Alhindawi N, Samara G, Abuasal S, Alazaidah R. A Deep Learning Approach Towards Student Performance Prediction in Online Courses: Challenges Based on a Global Perspective. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1-6.

64. Al-Batah MS. Integrating the principal component analysis with partial decision tree in microarray gene data. IJCSNS Int J Comput Sci Netw Secur. 2019;19(3):24-29.

65. Alazaidah R, Owida HA, Alshdaifat N, Issa A, Abuowaida S, Yousef N. A comprehensive analysis of eye diseases and medical data classification. TELKOMNIKA (Telecommunication Comput Electron Control. 2024;22(6):1422–30.

66. Al-Batah MS. Ranked features selection with MSBRG algorithm and rules classifiers for cervical cancer. Int J Online Biomed Eng. 2019;15(12):4.

67. Alazaidah R. A Comparative Analysis of Discretization Techniques in Machine Learning. In: 2023 24th International Arab Conference on Information Technology, ACIT 2023. 2023. p. 1–6.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.kaggle.com/datasets/saidulkabir/vcug-vur-dataset

FINANCING

This work is supported from Jadara University under grant number [Jadara-SR-Full2023], Zarqa University, and The University of Jordan.

CONFLICT OF INTEREST

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon, Esra Alzaghoul.

Research: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon, Esra Alzaghoul.

Methodology: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon, Esra Alzaghoul.

Writing - original draft: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon, Esra Alzaghoul.

Writing: review and proof editing: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon, Esra Alzaghoul.