JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 02, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0302-0070

Data Pitfalls in Explainable ML for UPJ Obstruction in Horseshoe Kidney

Dharmendra Chavda[1,*], Krushal Gohil[1], Harshwardhansinh Chauhan[1], Vinayak Pandya[1], Jaimin Chavda[2]

[1] Drs. Kiran & Pallavi Patel Global University (KPGU), Vadodara, Gujarat, India
[2] Sigma University, Vadodara, Gujarat, India

*Corresponding author: dharmendra.chavda.it@gmail.com

Contributing authors: gohilkrushal110@gmail.com, hwkc392@gmail.com, vinayakpandya2001@gmail.com, chavdajaimin.it@gmail.com

Received: October 16, 2026 **Accepted:** November 10, 2026

Abstract- Assessment of ureteropelvic junction obstruction (UPJO) in horseshoe kidney (HSK) is difficult, and curated data for this congenital anomaly are scarce, which makes synthetic data attractive for building machine-learning (ML) processes. How much can be concluded from such data is rarely evaluated. We report a reproducible methodological case study on a synthetic cohort of 1,000 HSK records (no real patients) and ask how far headline performance and explanation plots can be trusted. Post-treatment fields were excluded to prevent target leakage. On a balanced 80:20 holdout (N=200), XGBoost reached ROC-AUC 0.997 and accuracy 0.955, and Logistic Regression, SVM, and Random Forest performed comparably (ROC-AUC 0.994-0.996). In repeated cross-validation, all four models reached ROC-AUC 0.990-0.993, while logistic regression reached 0.983. When a random fraction of labels was flipped (10 independent draws), observed ROC-AUC fell to 0.93 at 5%, 0.87-0.88 at 10%, and 0.76-0.77 at 20%, whereas Gaussian measurement noise of $1.0 \times$ SD on APD_mm and $T_{1/2}$ reduced it only to 0.95-0.96. As separate process checks that are not evidence about HSK or UPJO, the same workflow reached ROC-AUC $\geq$ 0.999 on the easy public UCI Chronic Kidney Disease dataset (n = 400) but only 0.81-0.85 on a harder public urine-analysis dataset (n = 79). Near-perfect metrics on the synthetic cohort therefore reflect rule-based labels, not diagnostic certainty. We give reporting and stress-testing recommendations for ML studies that use synthetic clinical data. This is a methodological study, not a clinically validated diagnostic tool.

Keywords- Diuretic renography; explainable artificial intelligence (XAI); extreme gradient boosting (XGBoost); horseshoe kidney; label noise

1. Introduction

Horseshoe kidney (HSK) is the most common renal fusion anomaly, affecting about 1 in 400 to 1 in 800 people [1], [2]. Failure of axial rotation and arrested cephalad migration produce a fibrous or parenchymal isthmus joining the lower poles across the midline [1]. The renal pelves therefore lie anteriorly, the ureters cross the isthmus, ureteral insertions are often high, and anomalous lower-pole arteries arising directly from the aorta or iliac vessels can compress the collecting system [2], [3]. These features predispose patients to urinary stasis, recurrent nephrolithiasis and ureteropelvic junction obstruction (UPJO) [2], [3].

Assessing UPJO in this anomalous anatomy is difficult. A typical work-up combines ultrasound, including the Society for Fetal Urology (SFU) hydronephrosis grade and the maximal anteroposterior pelvic diameter (APD) [4-5], with diuretic renography (e.g., ^{99m}Tc-MAG3) to obtain differential renal function (DRF), drainage half-time ($T_{1/2}$) and the shape of the clearance curve are shown in Figure 1 and Figure 2. In HSK, however, delayed tracer washout is common and may reflect gravity-dependent pooling in anteriorly rotated calyces rather than true mechanical obstruction [2], [6]. Even in

anatomically normal kidneys, the decision to operate is made from several imperfect surrogates rather than a single gold standard [4-5]. Objective decision support that separates benign stasis from progressive obstruction needing pyeloplasty would therefore be valuable [6].

Machine-learning (ML) models such as extreme gradient boosting (XGBoost) [7] perform well on many tabular medical prediction tasks [8], but clinical translation is limited by their opacity and by the scarcity of data for rare congenital anomalies [9]. Explainable-AI (XAI) methods such as SHapley Additive exPlanations (SHAP) are often used to address the first problem [9], [10], and synthetic data are increasingly used to address the second [11], [12]. Synthetic tabular data allow processes to be built safely, but they need caution: artifacts of the generating process can inflate accuracy while omitting the difficulties of noisy biological cohorts [11], [13]. When both performance and explanations are computed on the same synthetic data, it is easy to mistake a model that has recovered the generator's rule for a model that has learned clinical physiology.

This study asks three methodological questions.

- RQ1: How much of the near-ceiling performance on a synthetic HSK-UPJO cohort is explained by a simple, low-dimensional rule?
- RQ2: What do SHAP explanations reveal on such data, and what can they not establish?
- RQ3: How sensitive are the headline metrics to realistic sources of error, namely reference-standard (label) disagreement and measurement noise on the key imaging variables?

We train an explainable workflow (XGBoost, Random Forest, support vector machine (SVM) and Logistic Regression, with SHAP [9], [10]) on the synthetic HSK-UPJO dataset and then stress-test it. Our contributions are: (i) a leakage-aware feature set with a documented audit of flawed fields in the source file; (ii) simple and rule-based baselines with repeated cross-validation that quantify how much signal the two dominant variables carry; (iii) label-noise and measurement-noise simulations, each repeated over 10 independent draws, that quantify how quickly the metrics degrade; (iv) supplementary runs on two public real-world datasets, one easy (UCI Chronic Kidney Disease) and one hard (urine analysis), which show that the same process produces near-ceiling or modest metrics depending on the task, and which are not presented as evidence about HSK or UPJO; and (v) release of the code, a corrected dataset and a single script that regenerates every reported number and figure. No real HSK patient is analyzed, and no clinical validity is claimed.

The rest of the paper is organized as follows. Section 2 reviews related work and states the research gap. Section 3 describes the dataset, the leakage controls, the models and the validation and stress-testing protocol. Section 4 reports the results, with the synthetic-cohort experiments (Sections 4.1-4.5) kept separate from the supplementary public-dataset experiments (Section 4.6). Section 5 discusses the findings against previous studies and their practical implications, Section 6 lists the limitations and Section 7 concludes.

2. Related Work

This section reviews the six bodies of work on which the study draws and ends with the gap it addresses. Table 1 summarizes representative ML studies of obstruction in hydronephrosis.

2.1. Machine Learning for UPJO and Obstructive Hydronephrosis

ML for obstruction has been studied mostly in children with antenatally detected hydronephrosis, and almost always with real data and an operational reference standard such as pyeloplasty or a renographic definition of obstruction. Blum et al. extracted features from the whole diuretic renogram curve and showed that a learned classifier detected clinically significant obstruction earlier and more accurately than the conventional $T_{1/2}$ and 30-min clearance [14]. Cerrolaza et al. derived quantitative ultrasound features of the collecting system as an imaging-only measure of obstructive severity [15]. Erdman et al. trained deep networks on ultrasound images alone to predict obstruction [16] and later developed the Hydronephrosis Severity Index, which was validated prospectively and at three external hospitals [17]. Khondker et al. used routine ultrasound parameters in a random-forest model (AERO), in which age, APD and SFU grade were the most important predictors, and reported an internal AUROC of 0.84 that fell to 0.76 on external validation [18], and later showed that adding ultrasounds from several visits did not improve on a single visit [19]. Drysdale et al. predicted recurrent obstruction after pyeloplasty in 543 children (AUROC 0.86) [20]. Alici and Tokar reported an XGBoost ROC-AUC of 0.977 for predicting surgery in 323 children with UPJO, with scintigraphic obstruction among the key predictors [21], and Wolmer et al. found that APD alone had an AUC of 0.867 for clinically significant events in a small prospective cohort [22]. None of these studies focuses on HSK, and none uses synthetic data. A recurring observation is that APD and renographic drainage carry most of the predictive signal, and that performance falls when models are tested on new institutions [17], [18].

2.2. Explainable ML and the Limits of Post Hoc Attribution

SHAP assigns each feature an additive contribution based on Shapley values [10], and TreeSHAP computes these exactly and efficiently for tree ensembles [9]; LIME is a widely used local alternative [23]. SHAP has been used to make clinical

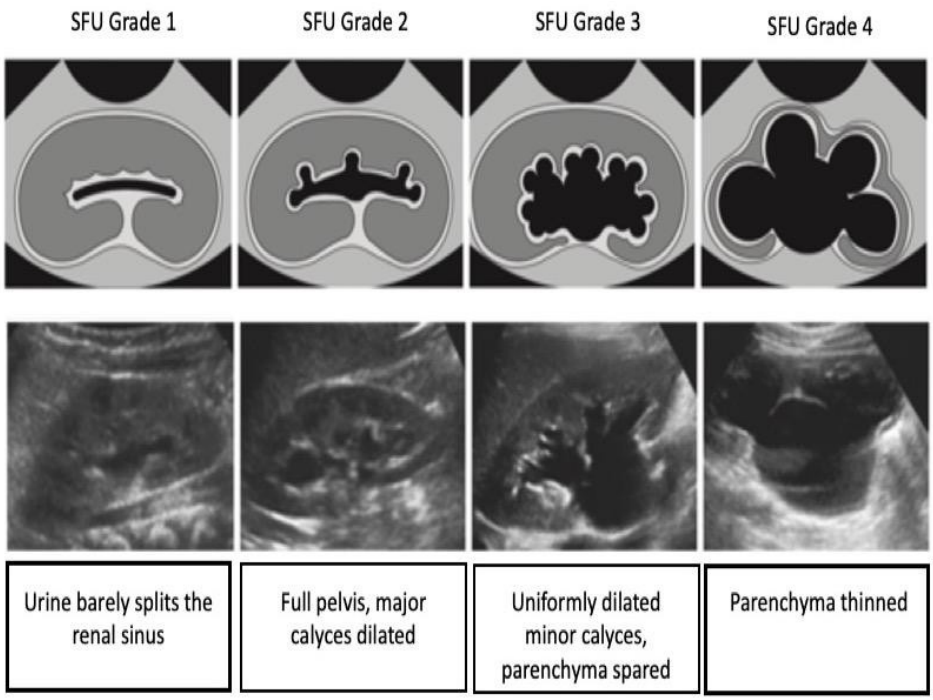


Figure 1: *Society for Fetal Urology (SFU) hydronephrosis grading. Grade 1: urine barely separates the renal sinus; Grade 2: full pelvis with dilated major calyces; Grade 3: uniformly dilated minor calyces with preserved parenchyma; Grade 4: dilation with parenchymal thinning (adapted from the SFU consensus statement [4]).*

risk models such as intra-operative hypoxaemia prediction inspectable by clinicians [24]. Its limits are equally well documented. Attributions explain the fitted model, not the data-generating process, and Shapley-based importance can be unstable or misleading when features are correlated or when the explanation game is poorly specified [25]. Post hoc explanations can even be manipulated to hide a model's reliance on a sensitive variable [26]. Ghassemi et al. argue that current XAI methods cannot by themselves establish that an individual clinical prediction is trustworthy [27], and Rudin recommends inherently interpretable models for high-stakes decisions [28]. These points matter more for synthetic data, where a model can reproduce the generator's rule exactly and SHAP will faithfully display that rule as if it were clinical knowledge.

2.3. Synthetic Clinical Data

Synthetic health data are produced by rule- or simulation-based generators such as Synthea [29] and by learned generators such as conditional GANs for tabular data [30] and privacy-oriented GANs [31]. Reviews emphasize their value for privacy, data sharing, augmentation of rare conditions and software testing, but also warn that fidelity, utility and privacy must

Table 1: *Representative Machine-Learning Studies of Obstruction in Hydronephrosis.*

Study	Data and Task	Reported Performance	Relevance to This Study
Blum *et al.* [14]	Real data; renogram-curve features; obstruction detection	Better than $T_{1/2}$ and 30-min clearance	Renogram signal dominates.
Cerrolaza *et al.* [15]	Real data; quantitative ultrasound; obstructive severity	Imaging-only severity measure	APD-type geometry is informative.
Erdman *et al.* [16], [17]	Real data; ultrasound images; surgery/obstruction prediction	High AUROC; externally validated in 2024	Demonstrates the importance of external validation.
Khondker *et al.* [18]	Real data from two sites ($n = 304$); routine ultrasound; obstruction prediction	AUROC 0.84 internally and 0.76 externally	Performance decreases on external data.
Drysdale *et al.* [20]	Real data ($n = 543$); re-intervention after pyeloplasty	AUROC 0.86	APD is a key predictor.
Alici and Tokar [21]	Real data ($n = 323$); prediction of surgery requirement in UPJO	XGBoost ROC-AUC 0.977	Near-ceiling performance on real clinical labels.
Wolmer *et al.* [22]	Real data ($n = 28$); clinically significant events	APD AUC 0.867	A simple APD-based baseline remains strong.
This study	Synthetic HSK data ($n = 1{,}000$); UPJO classification	ROC-AUC ≈ 0.99; two-feature baseline ROC-AUC 0.983	Stress-tests headline metrics and SHAP explanations.

Note: Performance values are reported as presented in the cited studies and are not directly comparable because of differences in datasets, tasks, validation protocols, and outcome definitions.

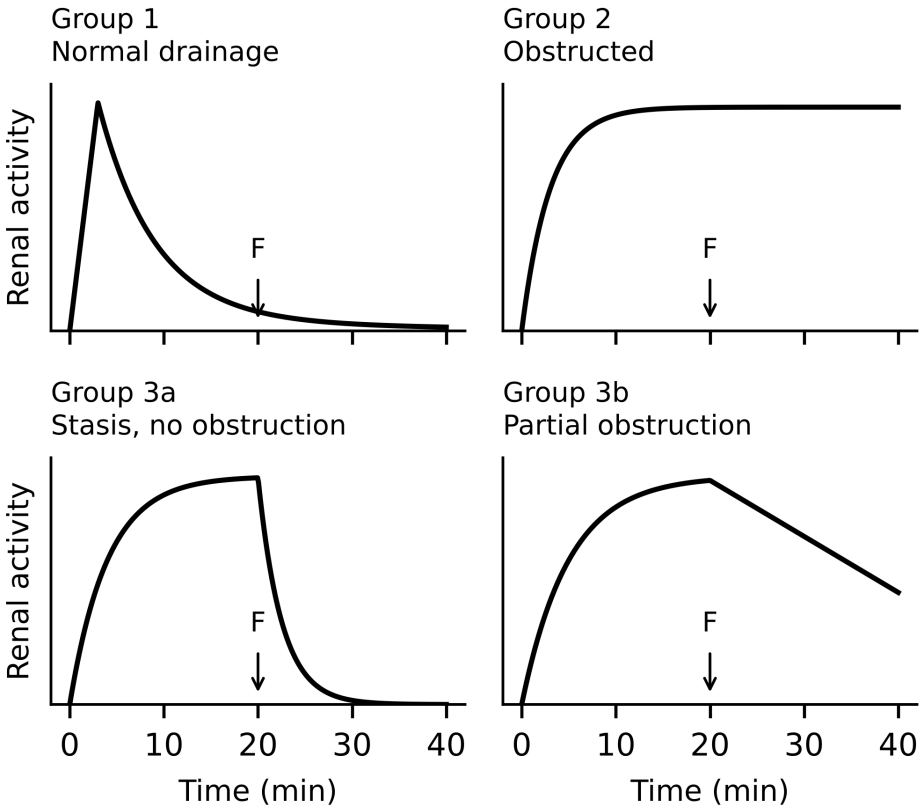


Figure 2: *Schematic diuretic renogram (^{99m}Tc-MAG3) curve typologies; F marks furosemide injection. Group 1: normal prompt drainage; Group 2: obstructed, with continued accumulation and no response to furosemide; Group 3a: non-obstructive stasis (hypotonic pelvis) with prompt response to furosemide; Group 3b: partial obstruction with slow, incomplete clearance (redrawn after the renography guidelines [5]). In this study only the drainage half-time $T_{1/2}$ derived from such curves is used as a predictor.*

be evaluated rather than assumed [11], [12], [32]. Alaa et al. propose sample-level fidelity, diversity and generalization metrics for auditing generators [33]. Most of this literature evaluates whether synthetic data resemble real data. Far less work asks the converse question addressed here: when a model is trained and evaluated on synthetic data only, which of its reported metrics and explanations carry information beyond the generator itself.

2.4. Label Noise and Reference-Standard Uncertainty

Classification with noisy labels is a mature field [34]. Confident learning estimates which labels are likely wrong [35], and errors in widely used test sets are frequent enough to change which model appears best [36]. In medical imaging, label noise from inter-observer disagreement and imperfect reference standards is pervasive [37]. UPJO is a clear case: there is no single gold standard, and the decision to operate depends on ultrasound grading, renographic interpretation and clinical course, each subject to variability [4], [5]. A synthetic label produced by a deterministic rule has none of this uncertainty, so robustness to label noise must be tested explicitly.

2.5. Target Leakage and Evaluation Pitfalls

Leakage occurs when information unavailable at prediction time, or derived from the outcome, enters the predictors [38]. Kapoor and Narayanan documented leakage across many fields of ML-based science and proposed a taxonomy that includes post-treatment variables and non-independence between training and test data [13]. Leakage inflated performance in neuroimaging [39] and genomics [40], and the review of COVID-19 imaging models by Roberts et al. found methodological flaws, including confounding and inappropriate evaluation, in nearly all the studies they examined [41]. Selecting models or tuning on the evaluation data also biases cross-validated estimates [42].

2.6. Robustness, Baselines and Reporting

For clinical tabular data, complex ML often offers no benefit over logistic regression [43], so simple baselines are an essential reference. Calibration, not only discrimination, should be reported [44]; models can fail under dataset shift

between sites [45]; and methodological failures in medical ML are often traced to small or unrepresentative evaluation data [46]. TRIPOD+AI [47], the urology-specific STREAM-URO framework [48] and the PROBAST risk-of-bias tool [49] set out what prediction-model studies should report, and sample-size guidance shows how many events a credible clinical model needs [50].

2.7. Research Gap and Novelty

Prior UPJO studies use real but small, mostly single-condition cohorts and report discrimination but none addresses HSK specifically. The synthetic-data literature concentrates on generating and auditing data, and the leakage and label-noise literature is largely outside urology. To our knowledge, no study has examined, in a rare congenital urological condition, what can and cannot be concluded when performance and explanations are both computed on synthetic data. The novelty of this work is therefore methodological: it combines leakage auditing, rule-recovering baselines, repeated label and measurement-noise stress tests and contrasting real-data controls into one reproducible protocol, and turns the results into concrete reporting recommendations (Section 5.5).

3. Materials and Methods

3.1. Study Design and Reporting

This is a methodological benchmarking study of a synthetic cohort. Reporting follows the applicable items of TRIPOD+AI [47] and STREAM-URO [48]; items that require real patients (e.g., participant flow, external validation) are not applicable. The synthetic-cohort experiments (Sections 3.2-3.8, results in Sections 4.1-4.5) and the supplementary public-dataset experiments (Section 3.9, results in Section 4.6) are kept separate throughout, because the latter concern different diseases and are used only to check the behavior of the process.

3.2. Dataset Provenance, Audit and Label-Generation Assumptions

The cohort consists of exactly 1,000 synthetic records with 18 columns, distributed as HSK_UPJO_synthetic_1000.csv. The file was produced by a data-generation script and contains no real patient information. A structural audit found no missing values, no duplicated rows and 1,000 unique Patient_ID values. The binary outcome UPJO is positive in 347 records (34.7%) and negative in 653 (65.3%) as shown in Figure 3.

Provenance has few limitations. The script that generated the data was not retained. We therefore cannot state the generator's equations, and we treat the label-generating structure as unknown and characterize it a posteriori from the data. Four observations constrain it. (i) The outcome is almost fully determined by APD_mm and Diuretic_renogram_T1_2_min: a two-feature logistic regression reaches ROC-AUC 0.983 and a depth-3 decision tree on the same two variables reaches accuracy 0.934 (Section 4.4). (ii) Variables that are clinically coupled are only weakly coupled in the file: APD_mm correlates with the SFU grade at Spearman $\rho = 0.22$, although the SFU grade is defined by the degree of pelvicalyceal dilation [4]. (iii) Bounded variables show hard limits consistent with clipping (APD_mm 2.0-30.0 mm; $T_{1/2} \geq 5.0$ min; DRF 21.8-55.0%). (iv) Surgical_modality and Postoperative_DRF_percent are populated for all 1,000 records, including the 653 without obstruction, which is clinically impossible and indicates that these fields were simulated without reference to the outcome (Section 3.3). Because the generator is unknown, the reproducibility claimed in this paper is that of the analysis: every number, table and data-driven figure is regenerated from the released file by one script (Section 3.10).

3.3. Feature Taxonomy and Prevention of Target Leakage

A clinical prediction tool must use only information available at the time of the decision; anything measured after treatment, or derived from the outcome, causes target leakage [13], [38]. The raw attributes cover demographics, clinical history, ultrasonography, diuretic renography, anatomical anomalies and surgery. Each was classified by clinical role and by when it becomes available as shown in Table 2. This classification determined its process status, as follows.

- Patient_ID was excluded because it is an arbitrary key with no clinical meaning.
- Postoperative_DRF_percent was excluded because it is measured months after surgery; using it would be direct target leakage.
- Surgical_modality (open, laparoscopic or robotic pyeloplasty) was excluded because it is chosen after the diagnosis and would introduce circular, post-treatment bias.

The remaining 14 baseline variables form the predictor set: Age_years, Sex, Laterality, Prenatal_hydronephrosis, Initial_presentation, Flank_pain, Recurrent_UTI, Hydronephrosis_SFU_grade, APD_mm, Preoperative_DRF_percent, Diuretic_renogram_T1_2_min, High_ureteral_insertion, Crossing_vessel and Renal_stone (Table 2). Preoperative_DRF_percent is available before surgery and was retained, but because DRF is itself part of the obstruction work-up [5] it was also removed in a sensitivity ablation (Section 4.2). In the source file, Surgical_modality and Postoperative_DRF_percent are complete for all 1,000 records, including the 653 non-obstructed ones; this is a data-quality defect of the generator, and

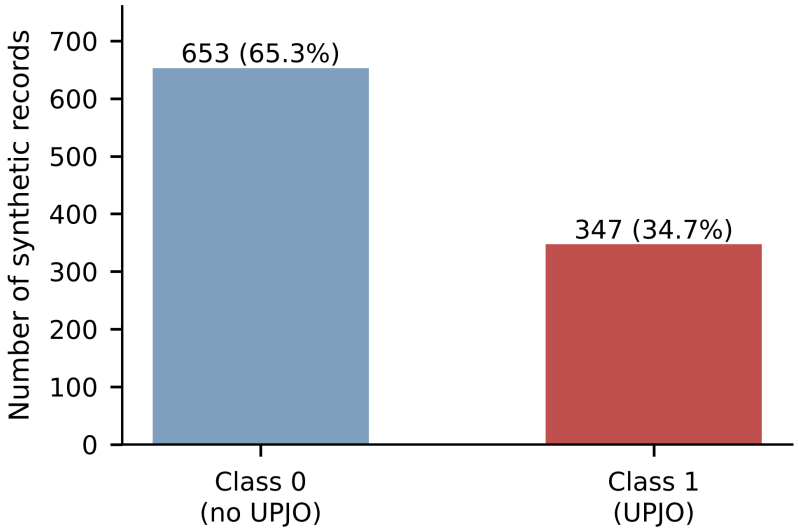


Figure 3: *Class distribution of the UPJO target in the synthetic horseshoe-kidney cohort (N = 1,000): Class 0 (no UPJO) 653 (65.3%), Class 1 (UPJO) 347 (34.7%).*

the corrected release (v2) leaves both fields blank when UPJO = 0. Neither field is used by any model, so the correction does not change any result.

Table 2: *Feature Taxonomy, Clinical Role, and Process Status.*

Variable	Type / Unit	Clinical Domain	Process Status
Patient_ID	ID	Administrative	Excluded (key)
Age_years	Continuous (yr)	Demographics	Predictor
Sex	Binary	Demographics	Predictor
Laterality	Binary (L/R)	Anatomical side	Predictor
Prenatal_hydronephrosis	Binary	History	Predictor
Initial_presentation	Categorical (4)	Presentation	Predictor
Flank_pain	Binary	Symptom	Predictor
Recurrent_UTI	Binary	Infection	Predictor
Hydronephrosis_SFU_grade	Ordinal (2-4)	Ultrasound	Predictor
APD_mm	Continuous (mm)	Ultrasound	Predictor
Preoperative_DRF_percent	Continuous (%)	Renography	Predictor; ablated
Diuretic_renogram_T1_2_min	Continuous (min)	Renography	Predictor
High_ureteral_insertion	Binary	HSK anatomy	Predictor
Crossing_vessel	Binary	Vascular anomaly	Predictor
Renal_stone	Binary	Complication	Predictor
Surgical_modality	Categorical (3)	Treatment	Excluded (post-treatment)
Postoperative_DRF_percent	Continuous (%)	Outcome after surgery	Excluded (leakage)
UPJO	Binary	Reference label	Target

3.4. Univariate Statistical Audit

Before modeling, baseline associations were examined with non-parametric tests. For continuous variables, the Mann-Whitney U test showed large differences between the non-obstructed (Class 0) and obstructed (Class 1) groups:

- APD_mm: median 8.4 mm vs. 18.8 mm (U = 10,970, $p < 0.001$).
- Diuretic_renogram_T1_2_min: median 13.4 min vs. 29.8 min (U = 14,479, $p < 0.001$).
- Preoperative_DRF_percent: median 46.1% vs. 36.7% (U = 201,654.5, $p < 0.001$).
- Age_years: median 12.1 vs. 14.2 years (U = 95,635.5, $p < 0.001$).

For categorical variables (χ^2 test), Crossing_vessel ($\chi^2 = 27.78$, $p < 0.001$) and High_ureteral_insertion ($\chi^2 = 44.66$, $p < 0.001$) were associated with UPJO, and the SFU grade only weakly so ($\chi^2 = 9.81$, $p = 0.007$; Figure 4). Flank_pain ($p = 0.085$), Recurrent_UTI ($p = 0.232$) and Renal_stone ($p = 0.958$) were not. These tests describe the synthetic file; they are reported because they foreshadow which variables the models will rely on, not as evidence of clinical association.

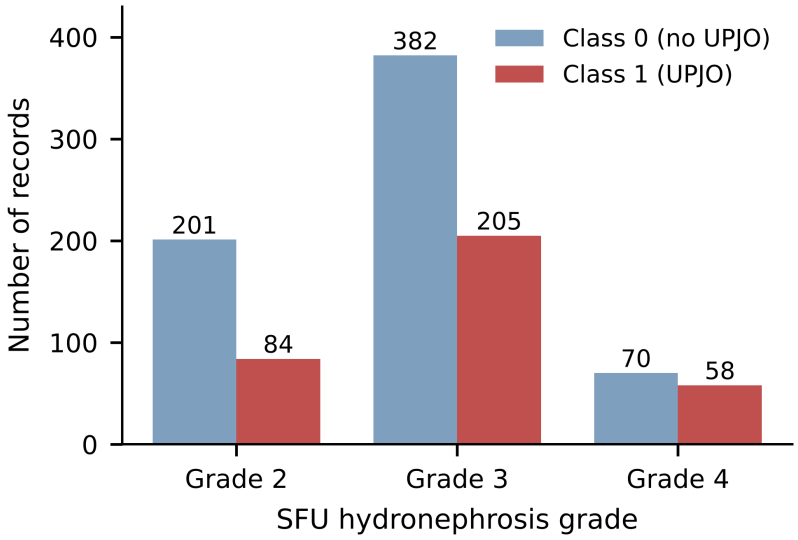


Figure 4: *SFU hydronephrosis grade by UPJO status in the synthetic cohort. Grade 2: 201 controls vs. 84 obstructed; Grade 3: 382 vs. 205; Grade 4: 70 vs. 58. The weak separation by grade, despite strong separation by APD, is one of the signs that the generator did not model clinical coupling between variables.*

Table 3: *Model Configurations for the Synthetic-Cohort and Urine-Analysis Experiments.*

Model	Configuration
Logistic Regression	L2 penalty, $C = 1.0$, `lbfgs` solver, `max_iter`=1000; standardized inputs.
SVM	RBF kernel, $C = 1.0$, $\gamma =$ `scale`, Platt-scaled probabilities; standardized inputs.
Random Forest	300 trees, `class_weight=balanced`; remaining parameters set to default values.
XGBoost	300 trees, `max_depth`=4, `learning_rate`=0.05, `subsample`=0.8, `colsample_bytree`=0.8; log-loss objective.
Baselines	Logistic Regression using APD_mm; Logistic Regression using $T_{1/2}$; Logistic Regression using APD_mm + $T_{1/2}$; and a depth-3 decision tree using APD_mm + $T_{1/2}$.

Note: Random seed 42 was used for all models. A decision threshold of 0.5 was used to compute accuracy, precision, recall, and F1-score.

3.5. Preprocessing and Model Configurations

The records were split into a balanced 80:20 training and holdout set (random_state = 42): 800 training cases (522 Class 0, 278 Class 1) and 200 holdout cases (131 Class 0, 69 Class 1). Preprocessing was implemented in a scikit-learn ColumnTransformer inside each model process, so that it was fitted on training data only. The nine categorical variables were one-hot encoded (drop = "first", handle_unknown = "ignore"), giving 11 indicator columns; the five numeric variables (Age_years, Hydronephrosis_SFU_grade, APD_mm, Preoperative_DRF_percent and Diuretic_renogram_T1_2_min) were passed through unchanged for the tree models and standardized (fitted on the training fold) for Logistic Regression and SVM, giving $M = 16$ model inputs. Four classifiers were trained with fixed hyperparameters as presented in Table 3. No hyperparameter search was performed, so no tuning information can leak from the evaluation data [42]. The same configuration is used in every experiment (holdout, repeated cross-validation and noise tests).

3.6. Validation Procedures

Three complementary procedures were used:

- Holdout: metrics on the 200 holdout cases (accuracy, precision, recall, F1, ROC-AUC and the Brier score), with 95% confidence intervals from 2,000 bootstrap resamples of the holdout set [51].
- Repeated cross-validation: 5-fold cross-validation repeated 3 times with different shuffles (15 folds in total, seeds 42-44), reported as mean ± SD across folds. This is the primary estimate because a single 200-case split is sensitive to which cases it contains.
- Stress tests: the noise experiments of Section 3.8. All preprocessing is refitted inside every training fold.

3.7. Model Explainability (SHAP)

Predictions of the XGBoost model were explained with TreeSHAP [9], [10], which decomposes the model output (log-odds) for a case x into additive feature contributions:

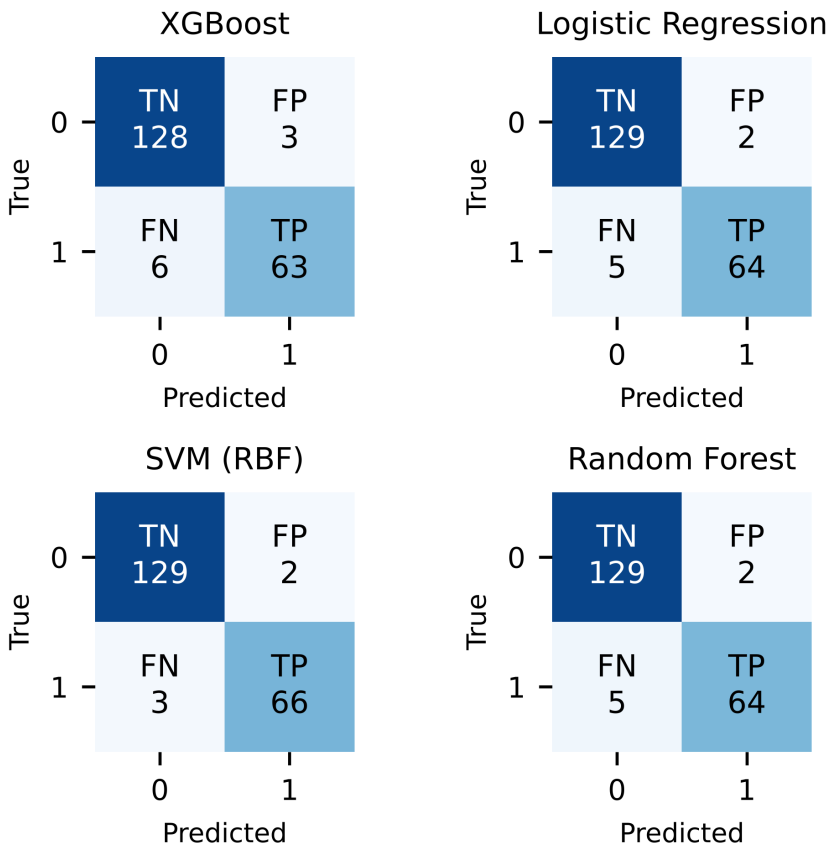


Figure 5: *Confusion matrices of the four classifiers on the 200-case holdout set.*

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \tag{1}$$

where $M = 16$ is the number of model inputs, ϕ_0 is the expected model output over the training data and $\phi_i(x)$ is the contribution of input i. SHAP values were computed for the 200 holdout cases. They measure contributions within the learned model and do not indicate clinical causality [25], [27].

3.8. Baselines and Noise Experiments

To quantify how much signal the two dominant measurements carry, univariate logistic regressions on APD_mm and on Diuretic_renogram_T1_2_min, a two-variable logistic regression and a depth-3 decision tree on both variables were evaluated with the same repeated cross-validation. Robustness was assessed in two ways.

1 *Label noise.* A random fraction (5%, 10%, 20% or 30%) of UPJO labels was flipped. In the observed-label scenario the flipped labels were used for both training and evaluation, which corresponds to a cohort whose reference standard is imperfect. In the training-only scenario models were trained on flipped labels and evaluated against the original labels.

2 *Measurement noise.* Zero-mean Gaussian noise with SD equal to 0.25, 0.5 or 1.0 times the empirical SD of the feature was added to APD_mm and Diuretic_renogram_T1_2_min, to represent inter-observer and hydration-related variability.

Logistic Regression, Random Forest and XGBoost were used. Each setting was repeated for 10 independent random draws; for each draw a fresh noise pattern was generated and the models were evaluated with balanced 5-fold cross-validation (seed 42 + draw index). Results are reported as the mean ± SD of the 10 draw-level ROC-AUC values. Random flipping is a simplification of real diagnostic disagreement (Section 6).

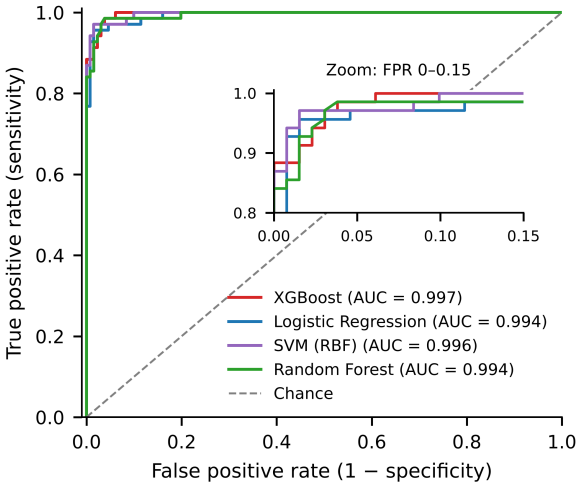


Figure 6: *ROC curves of the four classifiers on the holdout set; the inset magnifies false-positive rates of 0-0.15. The near-overlap of the curves reflects the rule-based synthetic labels rather than a clinically meaningful difference between models.*

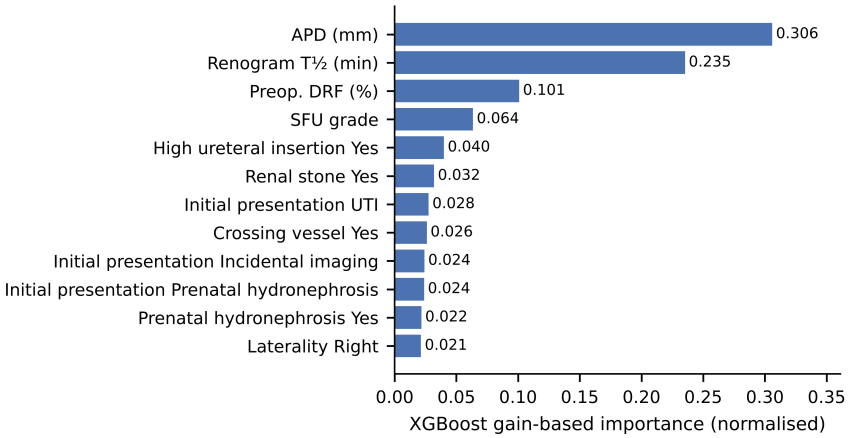


Figure 7: *Gain-based feature importance of the XGBoost model. Pelvic dilation (APD) and renographic drainage half-time ($T_{1/2}$) dominate.*

3.9. Supplementary Experiments on Public Real-World Datasets

These experiments test the behavior of the process on real measurements. They involve different diseases and are not evidence of validity for UPJO or HSK. First, we used the UCI Chronic Kidney Disease dataset [52], which contains 400 patients (250 with CKD) recorded at Apollo Hospitals, Tamil Nadu, India, with 24 clinical and laboratory attributes and 10.5% missing cells, as an easy real task. Second, as a real urological task of genuine difficulty, we used the urine-analysis dataset of Andrews and Herzberg [53], distributed with the boot package for R [51] and the Rdatasets collection. It contains 79 urine specimens (34 with and 45 without calcium oxalate crystals) and six physicochemical predictors: specific gravity, pH, osmolarity, conductivity, urea concentration and calcium concentration (two values missing; 0.4% of cells). The four classifiers and hyperparameters of Table 3 were evaluated by balanced 5-fold cross-validation repeated 5 times, with median imputation fitted within each training fold and standardization for Logistic Regression and SVM. A 2,000-sample bootstrap 95% CI of the ROC-AUC was computed on the out-of-fold predictions averaged over the five repeats. Single-predictor logistic regressions served as baselines, and TreeSHAP attributions were computed for an XGBoost model fitted to all specimens. The CKD analysis (ckd_pipeline.py) uses the same protocol and model families with minor configuration differences recorded in the script (e.g., 200 XGBoost trees of depth 3, and most-frequent imputation and one-hot encoding of nominal attributes). The urine dataset concerns calcium oxalate crystal formation, not HSK or UPJO.

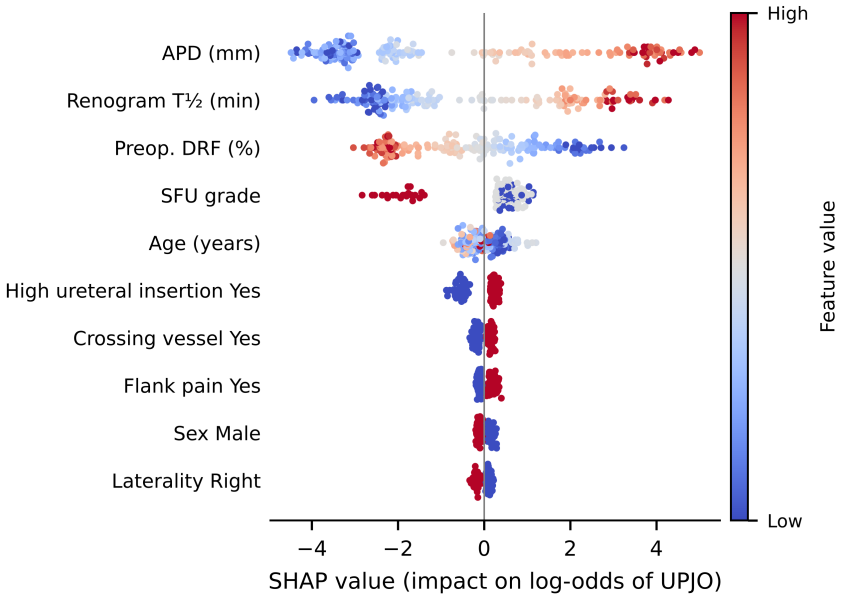


Figure 8: *TreeSHAP beeswarm summary for the XGBoost model on the holdout set (top 10 inputs). The horizontal axis is the contribution to the log-odds of UPJO and the color encodes the feature value (red high, blue low).*

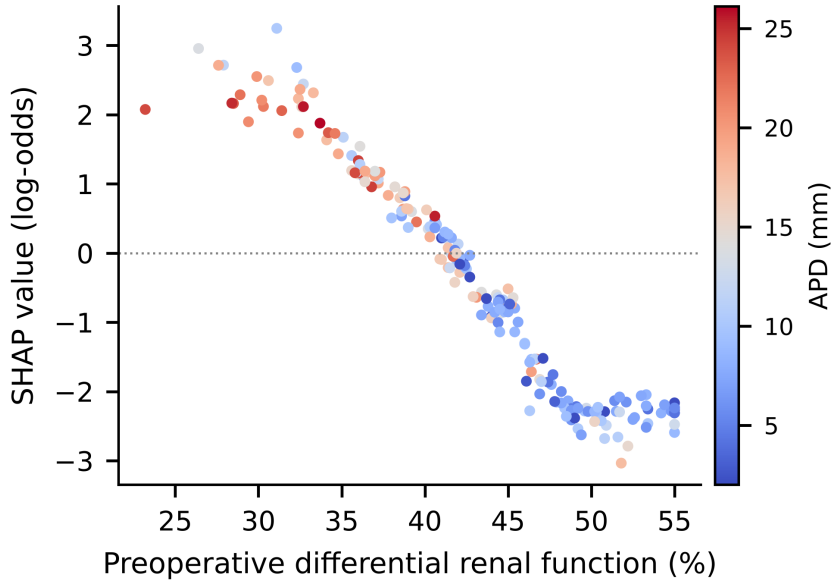


Figure 11: *SHAP dependence plot for Preoperative_DRF_percent, colored by APD_mm. Attributions become negative as split function rises above about 45% (model behavior, not a causal effect).*

3.10. Reproducibility

All synthetic-cohort and urine-analysis results were regenerated for this revision by one script (reproduce_all.py) from the released files, with Python 3.11, scikit-learn 1.9.1, XGBoost 3.4.1, SHAP 0.54.1, pandas 3.0.5 and NumPy 2.5.3, and seed 42 throughout; the CKD results are produced by ckd_pipeline.py from the UCI file. The script writes every table as a CSV file and every data-driven figure at 600 dpi. On a two-core machine the full run takes about 15 min, most of it for the 10-draw noise experiments. The original exploratory notebook is also released; because it was run with earlier library versions and without standardization for Logistic Regression and SVM, its holdout figures differ from those reported here by at most three test cases per model.

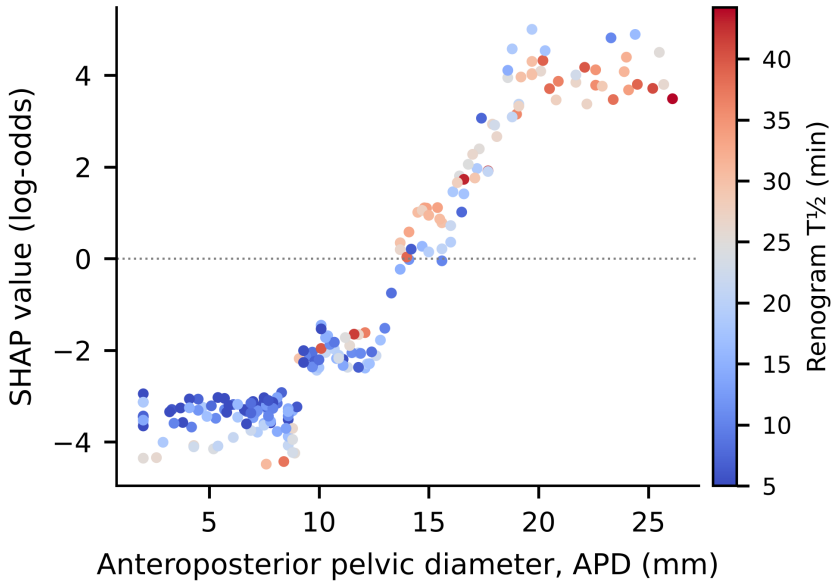


Figure 9: *SHAP dependence plot for APD_mm, colored by $T_{1/2}$. The steep rise between about 12 and 17 mm separates the two synthetic classes and reflects the label-generating rule.*

4. Results

4.1. Synthetic Cohort: Multi-Model Holdout Benchmark

On the same 200-case holdout set as in Table 4, linear and ensemble classifiers performed comparably. All four models reached ROC-AUC between 0.994 and 0.997, and their bootstrap 95% CIs overlap almost completely, so the holdout split cannot rank them. SVM had the highest point estimates on this split (accuracy 0.975, F1 0.964), whereas XGBoost had a ROC-AUC of 0.997 (95% CI 0.991-0.999) but a slightly lower accuracy of 0.955 (95% CI 0.925-0.980). The confusion matrices (Figure 5) show that the models differ by only a few cases: XGBoost made 3 false-positive and 6 false-negative errors, Logistic Regression and Random Forest 2 and 5, and SVM 2 and 3. A difference of one or two cases changes accuracy by 0.005-0.01, which is within the fold-to-fold SD of the repeated cross-validation as shown in Table 5. Figure 6 shows ROC curves that are almost indistinguishable and reach a sensitivity of about 0.9 or more at a false-positive rate of 0.05.

Brier scores were low (0.021-0.031), and out-of-fold Brier scores in repeated cross-validation were 0.028-0.036. Good calibration is expected here because the models are trained and tested on labels produced by the same rule; it is not evidence of clinical calibration [44].

Table 4: *Holdout Performance on the Synthetic Cohort (N = 200).*

Model	Acc.	Prec.	Rec.	F1	AUC	95% CI	Brier
XGBoost	0.955	0.955	0.913	0.933	0.997	0.991-0.999	0.027
LR	0.965	0.970	0.928	0.948	0.994	0.986-0.999	0.024
SVM (RBF)	0.975	0.971	0.957	0.964	0.996	0.991-1.000	0.021
RF	0.965	0.970	0.928	0.948	0.994	0.986-0.999	0.031
XGBoost, no DRF	0.955	0.955	0.913	0.933	0.993	0.983-0.999	0.032

Note: The holdout set contains 131 controls and 69 obstructed cases. LR = logistic regression; RF = random forest (balanced); Acc. = accuracy; Prec. = precision; Rec. = recall (sensitivity); AUC = ROC-AUC. Metrics were computed at a decision threshold of 0.5. The 95% confidence intervals were estimated using 2,000 bootstrap resamples of the holdout set.

4.2. Feature Importance and Sensitivity Ablation

Gain-based importances from the trained XGBoost model (Figure 7) ranked APD_mm first (0.306), followed by Diuretic_renogram_T1_2_min (0.235) and Preoperative_DRF_percent (0.101); together the two imaging variables account for 54% of the total gain, and no input other than DRF exceeds 0.064. When Preoperative_DRF_percent was removed, XGBoost still achieved accuracy 0.955, precision 0.955, recall 0.913, F1 0.933 and ROC-AUC 0.993. Removing DRF left the holdout confusion matrix unchanged and lowered ROC-AUC only from 0.997 to 0.993, so DRF is largely redundant once

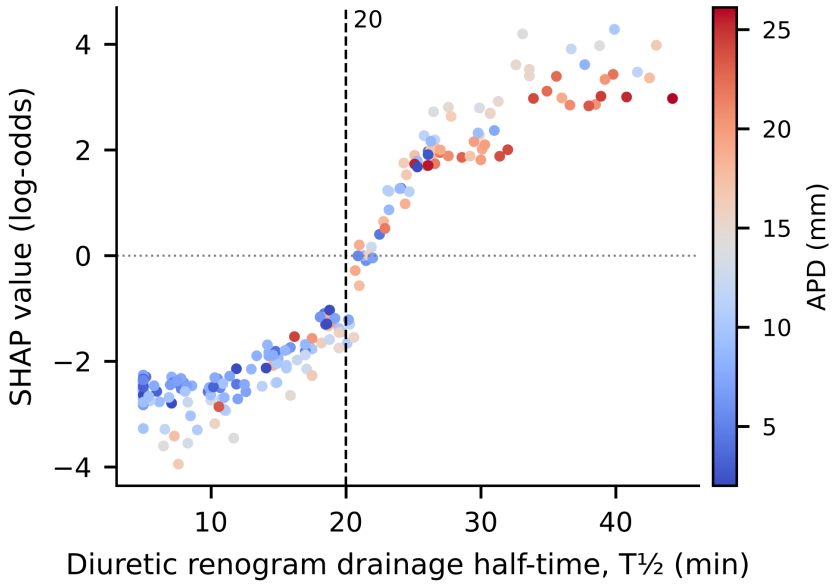


Figure 10: *SHAP dependence plot for Diuretic_renogram_T1_2_min, colored by APD_mm. The transition near 20 min (dashed line) coincides with the conventional delayed-drainage cut-off [5] but, in synthetic data, reflects the generator rule rather than clinical confirmation.*

APD_mm and $T_{1/2}$ are known. On real data this ablation would be clinically interesting, because DRF is not always available or reliable in HSK; on synthetic data it mainly confirms that the label depends on the two imaging variables.

4.3. Post Hoc Explainability with TreeSHAP

Figure 8 SHAP beeswarm plot shows global attributions: each point is one holdout case, the horizontal position is the SHAP value (contribution to the log-odds of UPJO) and the color is the feature value (red high, blue low). Features are ordered by mean |SHAP|. APD_mm (mean |SHAP| 2.82) and $T_{1/2}$ (2.20) dominate, followed by DRF (1.41); high APD and long $T_{1/2}$ push predictions toward UPJO, and preserved DRF pushes them away.

The dependence plots reveal threshold-like behavior. For APD_mm as shown Figure 9, attributions are strongly negative below about 10 mm, cross zero at about 13-14 mm (close to the first split of the depth-3 tree, 13.7 mm) and level off above about 18 mm. For $T_{1/2}$ as per Figure 10, the transition lies between about 20 and 22 min, which coincides numerically with the conventional 20-min cut-off for delayed drainage [5]. For DRF as shown in Figure 11, preserved function (> 45%) is attributed negatively and declining function (< 35%) positively.

4.4. Baselines and Repeated Cross-Validation

Repeated balanced 5-fold cross-validation gave ROC-AUC values of 0.990-0.993 for all four models with the 14 predictors; Logistic Regression was statistically indistinguishable from XGBoost. The baselines answer RQ1. A logistic regression on APD_mm alone reached ROC-AUC 0.951, on $T_{1/2}$ alone 0.936, and on the two together 0.983 (accuracy 0.939); a depth-3 decision tree on the same two variables reached accuracy 0.934. The other 12 predictors therefore add only about 0.011 to the ROC-AUC. A cohort whose label is almost fully determined by two thresholded measurements explains why a linear model equals a boosted ensemble [43] and why every model appears near-perfect. The tree's splits place the decision boundary at APD_mm $\approx$ 13.7 mm, with a $T_{1/2}$ threshold of roughly 22-27 min that depends on APD, and with APD_mm above about 18 mm sufficient on its own; this is our best a posteriori description of the generator's rule.

4.5. Robustness to Label and Measurement Noise

Table 6 and Figure 12(a) show the effect of flipping a random fraction of labels in both training and evaluation data (mean ± SD over 10 draws). ROC-AUC fell from about 0.99 to 0.93 at 5%, 0.87-0.88 at 10%, 0.76-0.77 at 20% and 0.64-0.67 at 30% label noise. The three models degrade almost identically, which shows that the loss is a property of the labels, not of the learning algorithm. The size of the drop is close to what would be expected if the models remained correct and only the evaluation labels were wrong. If a fraction ε of labels is flipped at random, a perfect ranker evaluated against the flipped labels has an expected observed ROC-AUC of

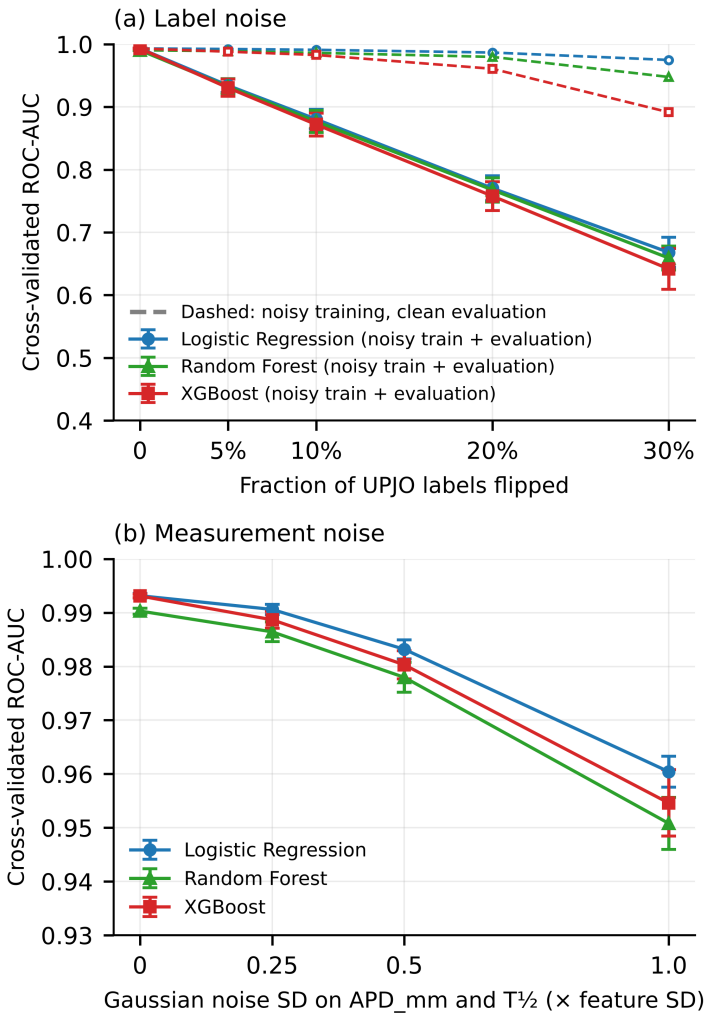


Figure 12: *Sensitivity of cross-validated ROC-AUC (mean ± SD over 10 independent draws) to (a) label noise, with labels flipped in training and evaluation data (solid) or in training data only (dashed), and (b) Gaussian measurement noise on APD_mm and $T_{1/2}$. Note the different vertical scales.*

$$\mathrm{AUC}_{\max}(\varepsilon) = \frac{A(\varepsilon)}{B(\varepsilon)} \tag{2}$$

where $A(\varepsilon) = (1-\varepsilon)^2 n_1 n_0 + \frac{1}{2}\varepsilon(1-\varepsilon)(n_1^2 + n_0^2)$, $B(\varepsilon) = [(1-\varepsilon)n_1 + \varepsilon n_0][(1-\varepsilon)n_0 + \varepsilon n_1]$, and n_1 and n_0 are the true class sizes (ties between cases of the same true class are counted as one half). For this cohort, (2) gives 0.941, 0.886, 0.781 and 0.684 at 5%, 10%, 20% and 30%; the observed values lie just below this ceiling, so almost all of the loss is explained by errors in the evaluation labels. When noise was present only in the training labels and models were evaluated against the original labels (dashed lines in Figure 12(a)), ROC-AUC remained 0.96-0.99 at 20% and 0.89-0.97 at 30% noise. The models are therefore fairly robust to noisy training labels; what limits the observed performance is the reliability of the labels used for evaluation.

Measurement noise had a much smaller effect as shown in Table 7 and Figure 12(b). At 1.0 × the feature SD, which is larger than typical inter-observer variability, ROC-AUC was still 0.95-0.96, partly because correlated predictors such as DRF compensated. In other words, the headline metrics are far more sensitive to errors in the reference standard than to errors in the predictors, and a real cohort with even 10% reference-standard disagreement could not show the near-ceiling values observed here (RQ3).

Table 5: *Baseline and Repeated Cross-Validation Performance on the Synthetic Cohort.*

Setting / Model	ROC-AUC (Mean ± SD)	Accuracy (Mean)
APD_mm only (LR)	0.951 ± 0.014	0.868
$T_{1/2}$ only (LR)	0.936 ± 0.015	0.876
APD_mm + $T_{1/2}$ (LR)	0.983 ± 0.006	0.939
APD_mm + $T_{1/2}$ (depth-3 tree)	0.964 ± 0.013	0.934
All 14: Logistic Regression	0.993 ± 0.002	0.960
All 14: Random Forest	0.990 ± 0.004	0.956
All 14: SVM (RBF)	0.994 ± 0.003	0.964
All 14: XGBoost	0.993 ± 0.004	0.956

Note: Results are based on repeated five-fold cross-validation with three repeats (15 folds in total). LR = logistic regression; SD = standard deviation across the 15 folds.

Table 6: *Cross-Validated ROC-AUC Under Label Noise.*

Model	0%	5%	10%	20%	30%
LR	0.993 ± 0.000	0.934 ± 0.011	0.880 ± 0.016	0.770 ± 0.020	0.668 ± 0.024
RF	0.990 ± 0.001	0.932 ± 0.013	0.877 ± 0.018	0.768 ± 0.019	0.659 ± 0.019
XGBoost	0.993 ± 0.000	0.930 ± 0.014	0.872 ± 0.018	0.758 ± 0.023	0.642 ± 0.032

Note: Labels were flipped in both training and evaluation data. Values are reported as mean ± standard deviation over 10 draws. LR = logistic regression; RF = random forest. At 0% label noise, the 10 draws differ only in the cross-validation shuffle.

4.6. Supplementary Experiments on Public Datasets (Not Evidence for HSK or UPJO)

The results in this subsection concern chronic kidney disease and urinary crystal formation. They show how the same process behaves on an easy and on a hard real task; they say nothing about diagnostic performance for UPJO in horseshoe kidneys.

On the UCI CKD dataset, all four models achieved a cross-validated ROC-AUC of 0.999 or higher. The task is easy: a logistic regression on hemoglobin alone reached ROC-AUC 0.965 and on serum creatinine alone 0.921. When the seven strongest laboratory indicators (serum creatinine, blood urea, hemoglobin, packed cell volume, red cell count, specific gravity and albumin) were removed, ROC-AUC remained 0.971 (Logistic Regression), 0.991 (Random Forest), 0.979 (SVM) and 0.988 (XGBoost). TreeSHAP ranked hemoglobin, specific gravity, serum creatinine and albumin highest. Near-ceiling metrics are thus not unique to synthetic data: they also occur on real data when the task itself is easy, which is why a high score alone cannot tell a reader whether a process has learned something difficult.

On the urine-analysis dataset, the four models achieved cross-validated ROC-AUC values of only 0.81-0.85 (accuracy 0.75-0.80), with wide bootstrap intervals because of the small sample (n = 79). Single-predictor logistic regressions reached ROC-AUC 0.799 for calcium and 0.754 for specific gravity; urea (0.667), osmolarity (0.644) and pH (0.564) were weak, and conductivity (0.474) was no better than chance. TreeSHAP ranked calcium (mean |SHAP| 1.66) and specific gravity (1.22) highest. Unlike the synthetic cohort, where all models were near ceiling, the best model here (Logistic Regression) is only modestly better than its strongest single predictor, the intervals of all four models overlap, and the fold-to-fold SD is about 0.10. These estimates, based on 79 specimens, are imprecise and do not refer to horseshoe-kidney performance.

Table 7: *Performance on the UCI Chronic Kidney Disease Dataset.*

Model	Accuracy	ROC-AUC	95% CI (AUC)
Logistic Regression	0.9925 ± 0.0106	0.9999 ± 0.0003	0.9993-1.0000
Random Forest	0.9910 ± 0.0090	0.9997 ± 0.0005	0.9990-1.0000
SVM (RBF)	0.9980 ± 0.0046	1.0000 ± 0.0000	1.0000-1.0000
XGBoost	0.9875 ± 0.0122	0.9990 ± 0.0014	0.9978-0.9999

Note: The UCI Chronic Kidney Disease dataset contains $n = 400$ samples. Results are based on repeated balanced five-fold cross-validation with five repeats. Accuracy and ROC-AUC values are reported as mean ± standard deviation.

5. Discussion

5.1. What the Near-Ceiling Metrics Mean

A ROC-AUC of about 0.997 on synthetic data does not demonstrate validated diagnostic accuracy, and the analyses above show why. The two-variable logistic regression reaches ROC-AUC 0.983, Logistic Regression matches XGBoost on all 14 predictors, and a three-level tree recovers the boundary, which is the signature of labels produced by a simple thresholding rule. These rules could only be characterized after the fact because the generator script was not kept. The comparison with real-data studies is instructive. Khondker et al., using routine ultrasound parameters on real patients (with age, APD and SFU grade as the strongest predictors), reported an internal AUROC of 0.84 that fell to 0.76 externally [18]; Drysdale

et al. reported 0.86 for re-intervention [20]; and even the best image-based models, trained on thousands of ultrasounds, report AUROCs around 0.9 [17]. The one real-data study with a value close to ours, an XGBoost ROC-AUC of 0.977 [21], included scintigraphic obstruction among its predictors and is internal validation only, so part of its outcome information is already present in the inputs. Values of 0.99 on a rule-labeled cohort should therefore be read as an upper bound set by the generator, not as an estimate of what is achievable in patients.

5.2. Explanations Recover the Generator, Not Physiology

In the importance and SHAP analyses, APD_mm and $T_{1/2}$ dominate the other variables, which is consistent with pediatric urology practice [4], [5] and with real-data models in which APD and renographic drainage are the strongest predictors [14], [18], [20], [22]. This agreement is reassuring about the generator's face validity, but it is circular as evidence: the generator was presumably written to reflect the same guidelines, and a model that reproduces it will display the guidelines back. This is the synthetic-data version of the concern raised by Ghassemi et al. that explanations can create unwarranted confidence [27]. Conversely, clinically important couplings that the generator did not encode, such as that between APD and SFU grade (ρ = 0.22 here), are invisible to any explanation method. Likewise, the significance of high ureteral insertion and crossing vessels reflects the generator's design, although it is consistent with the known anatomy of renal fusion, in which anomalous crossing vessels are common [1], [3].

5.3. Label Reliability Versus Measurement Noise

Real UPJO cohorts are complicated by biological variability, patient hydration during diuretic renography and inter-observer variability in ultrasound grading and renogram interpretation [4], [6]. Our noise experiments separate two of these sources. The observed ROC-AUC drops to about 0.76-0.77 when 20% of labels are discordant, whereas measurement noise of a full SD on the two key variables costs only about 0.03-0.04. This asymmetry agrees with the label-noise literature, in which errors in evaluation labels distort benchmarks and model rankings more than noise in the inputs [34], [36], [37]. The practical implication is that in a real HSK-UPJO cohort the dominant uncertainty will lie in the reference standard (the decision that a kidney is obstructed), and studies should invest in adjudicated labels and report inter-rater agreement. The contrast with the urine-analysis dataset, where the same models achieve only 0.81-0.85 (Section 4.6), shows that the process is not inherently near-perfect; its output depends on the task and the sample size, and results from a synthetic cohort cannot be generalized to real patients [13].

5.4. Prevention of Target Leakage

The process excludes the post-treatment variables Postoperative_DRF_percent and Surgical_modality, which were also implausibly populated for non-obstructed records. Their exclusion removes the most obvious leakage path and treatment confounding [13], [38] and restricts the predictors to preoperative information. Exclusion of post-treatment fields is necessary but not sufficient: because the labels are rule-based, the preoperative measurements APD_mm, $T_{1/2}$ and possibly DRF may have been used directly to construct the label. This is a form of circularity that cannot be checked without the generator, and it is the main reason why the reported accuracy is not an estimate of accuracy in practice.

5.5. Practical Implications and Recommendations

For researchers who use synthetic clinical data to develop or benchmark ML processes, our results support six recommendations: (1) release the generator script together with the data, or state clearly that it is unavailable; (2) audit fields for leakage and for clinically impossible values before modeling; (3) report the simplest rule-based or few-feature baseline alongside complex models; (4) report performance under label noise and measurement noise, with repeated draws; (5) present SHAP and importance plots as descriptions of the model and generator, not as clinical evidence; and (6) where possible, contrast the process with a real dataset of known difficulty. For clinicians and reviewers, a near-perfect AUC from synthetic data should prompt the question of which simple rule produced the labels. For HSK specifically, the next step is a multi-institutional registry with adjudicated obstruction labels, sample sizes chosen with established methods [50], and reporting according to TRIPOD+AI and STREAM-URO [47], [48].

6. Limitations

The principal limitations of this study are as follows.

- Only methodological conclusions can be drawn. Validation on clinical HSK registries, for example through pediatric urology anomaly consortia, is required before any clinical use.
- The data-generation script was lost, so the label rules were characterized only a posteriori (Sections 3.2 and 4.4), and circular use of preoperative measurements in label construction cannot be ruled out. We mitigated this by releasing the data, a corrected v2 file and a single script that reproduces every analysis, but the data themselves cannot be regenerated.

- Non-obstructed records contain Surgical_modality and Postoperative_DRF_percent values, which is clinically impossible. Both fields are excluded from all models, and the corrected release (v2) leaves them blank for non-obstructed records.
- No hyperparameter search and no external validation set were used. Random label flipping is a simplified model of real diagnostic disagreement, which is structured (for example, concentrated near decision thresholds). The noise experiments (10 draws per level) are therefore only an approximation.
- TreeSHAP values describe the behavior of the trained model, not medical or biological etiology, and thresholds in the dependence plots are properties of the synthetic labels.
- The CKD and urine-analysis experiments involve chronic kidney disease and calcium oxalate crystal formation (the latter with only 79 specimens and imprecise estimates). They show how the process behaves on real data and do not validate performance for horseshoe kidney.
- Results were regenerated with current library versions. The original exploratory notebook gives holdout results that differ by at most three test cases per model (Section 3.10).

7. Conclusion

This study presented a reproducible, explainable machine-learning workflow (XGBoost, Random Forest, SVM, Logistic Regression and TreeSHAP) for UPJO classification in a synthetic cohort of horseshoe kidneys, and used it to show why results from such data must not be read as clinical evidence. The models reached near-ceiling discrimination, but a two-variable logistic regression came within about 0.01 of them (RQ1); SHAP recovered threshold-like rules that mirror clinical conventions because the generator encoded them, not because the model learned physiology (RQ2), and observed performance fell sharply under label noise while remaining stable under measurement noise (RQ3). The same process was near-perfect on an easy public CKD dataset but reached only 0.81-0.85 on a harder public urine-analysis dataset, which confirms that such metrics depend on the task. The contribution of this work is methodological rather than clinical: a protocol of leakage audit, rule-recovering baselines, noise stress tests and real-data contrasts, together with reporting recommendations. Researchers who use synthetic clinical data should release the data and generator, report simple baselines and noise-sensitivity analyses and avoid clinical interpretation. Validation on multi-institutional HSK-UPJO registries remains essential before any translational claim.

Author Declarations

Author Contributions (CRediT): Dharmendra Chavda: Conceptualization; Methodology; Software; Formal analysis; Data curation; Visualization; Writing – original draft. Krushal Gohil: Investigation; Validation; Writing – review & editing. Harshwardhansinh Chauhan: Conceptualization; Supervision; Project administration; Writing – review & editing. Vinayak Pandya: Investigation; Visualization; Writing – review & editing. Jaimin Chavda: Validation; Resources; Writing – review & editing. All authors have read and approved the submitted version of the manuscript.

Funding: This study received no external funding and used no institutional resources or patient registries.

Data Availability Statement: The synthetic data (HSK_UPJO_synthetic_1000.csv and the corrected HSK_UPJO_synthetic_1000_v2.csv), the analysis scripts (reproduce_all.py, hsk_noise_experiments.py and ckd_pipeline.py), the original notebook and the result tables are available at https://github.com/ChavdaDharmendra/hsk-upjo-benchmark. A versioned archive of the revised repository is deposited at Zenodo: https://doi.org/10.5281/zenodo.23686317. The UCI Chronic Kidney Disease dataset is available from the UCI Machine Learning Repository [52], and the urine-analysis data from the boot R package and the Rdatasets collection [51], [53].

Acknowledgments: The authors thank the open-source machine-learning and scientific-computing communities (scikit-learn, XGBoost, SHAP and Matplotlib) whose software made this research possible.

Conflicts of Interest: The authors declares no conflict of interest.

Ethics Statement: This study used a synthetic dataset containing no real patients and two public, de-identified datasets [52], [53]. No human participants were recruited, and neither ethical approval nor informed consent was required.

Declaration of Generative AI: During the preparation of this work the authors used Grammarly to check grammar, spelling and language.

References

[1] K. Taghavi, J. Kirkpatrick, and S. A. Mirjalili, "The horseshoe kidney: Surgical anatomy and embryology," J. Pediatr. Urol., vol. 12, no. 5, pp. 275–280, Oct. 2016, doi: 10.1016/j.jpurol.2016.04.033.

[2] M. I. A. Elmaadawy, S. W. Kim, S. K. Kang, S. W. Han, and Y. S. Lee, "A retrospective analysis of ureteropelvic junction obstructions in patients with horseshoe kidney," Transl. Androl. Urol., vol. 10, no. 11, pp. 4173–4180, Nov. 2021, doi: 10.21037/tau-21-471.

[3] A. W. Partin, R. R. Dmochowski, L. R. Kavoussi, C. A. Peters, and A. J. Wein, Eds., Campbell-Walsh-Wein Urology, 12th ed. Philadelphia, PA, USA: Elsevier, 2020.

[4] H. T. Nguyen et al., "The Society for Fetal Urology consensus statement on the evaluation and management of antenatal hydronephrosis," J. Pediatr. Urol., vol. 6, no. 3, pp. 212–231, Jun. 2010, doi: 10.1016/j.jpurol.2010.02.205.

[5] I. Gordon, A. Piepsz, and R. Sixt, "Guidelines for standard and diuretic renogram in children," Eur. J. Nucl. Med. Mol. Imaging, vol. 38, no. 6, pp. 1175–1188, Jun. 2011, doi: 10.1007/s00259-011-1811-3.

[6] J. H. Ross and R. Kay, "Ureteropelvic junction obstruction in anomalous kidneys," Urol. Clin. North Am., vol. 25, no. 2, pp. 219–225, May 1998, doi: 10.1016/S0094-0143(05)70010-0.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD '16), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[8] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, Datasets and Benchmarks Track, 2022, pp. 507–520, doi: 10.52202/068431-0037.

[9] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Mach. Intell., vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[11] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," Nature Biomed. Eng., vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/s41551-021-00751-8.

[12] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," PLOS Digit. Health, vol. 2, no. 1, Art. no. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[13] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[14] E. S. Blum et al., "Early detection of ureteropelvic junction obstruction using signal analysis and machine learning: A dynamic solution to a dynamic problem," J. Urol., vol. 199, no. 3, pp. 847–852, Mar. 2018, doi: 10.1016/j.juro.2017.09.147.

[15] J. J. Cerrolaza et al., "Quantitative ultrasound for measuring obstructive severity in children with hydronephrosis," J. Urol., vol. 195, no. 4, pt. 1, pp. 1093–1099, Apr. 2016, doi: 10.1016/j.juro.2015.10.173.

[16] L. Erdman et al., "Predicting obstructive hydronephrosis based on ultrasound alone," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2020, Lecture Notes in Computer Science, vol. 12263. Cham, Switzerland: Springer, 2020, pp. 493–503, doi: 10.1007/978-3-030-59716-0_47.

[17] L. Erdman et al., "The Hydronephrosis Severity Index guides paediatric antenatal hydronephrosis management based on artificial intelligence applied to ultrasound images alone," Sci. Rep., vol. 14, no. 1, Art. no. 22748, Oct. 2024, doi: 10.1038/s41598-024-72271-9.

[18] A. Khondker et al., "Predicting obstruction risk using common ultrasonography parameters in paediatric hydronephrosis with machine learning," BJU Int., vol. 133, no. 1, pp. 79–86, Jan. 2024, doi: 10.1111/bju.16159.

[19] A. Khondker et al., "Longitudinal image-based prediction of surgical intervention in infants with hydronephrosis using deep learning: Is a single ultrasound enough?," PLOS Digit. Health, vol. 4, no. 8, Art. no. e0000939, Aug. 2025, doi: 10.1371/journal.pdig.0000939.

[20] E. Drysdale et al., "Personalized application of machine learning algorithms to identify pediatric patients at risk for recurrent ureteropelvic junction obstruction after dismembered pyeloplasty," World J. Urol., vol. 40, no. 2, pp. 593–599, Feb. 2022, doi: 10.1007/s00345-021-03879-z.

[21] C. A. Alici and B. Tokar, "Prediction of surgical necessity in children with ureteropelvic junction obstruction using machine learning," Ir. J. Med. Sci., vol. 194, no. 2, pp. 583–590, Apr. 2025, doi: 10.1007/s11845-025-03895-7.

[22] C. Wolmer, J. Delmas, S. Pecorelli, E. Dobremez, C. Ferdynus, and L. Harper, "Predicting clinically significant events in children with ureteropelvic junction obstruction," Front. Pediatr., vol. 12, Art. no. 1409170, 2024, doi: 10.3389/fped.2024.1409170.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD '16), San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[24] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," Nature Biomed. Eng., vol. 2, no. 10, pp. 749–760, Oct. 2018, doi: 10.1038/s41551-018-0304-0.

[25] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. A. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in Proc. 37th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 119, 2020, pp. 5491–5500.

[26] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES), New York, NY, USA, 2020, pp. 180–186, doi: 10.1145/3375627.3375830.

[27] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit. Health, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Mach. Intell., vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[29] J. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," J. Amer. Med. Inform. Assoc., vol. 25, no. 3, pp. 230–238, Mar. 2018, doi: 10.1093/jamia/ocx079.

[30] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019, pp. 7335–7345.

[31] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization through data synthesis using generative adversarial networks (ADS-GAN)," IEEE J. Biomed. Health Inform., vol. 24, no. 8, pp. 2378–2388, Aug. 2020, doi: 10.1109/JBHI.2020.2980262.

[32] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy," npj Digit. Med., vol. 6, no. 1, Art. no. 186, Oct. 2023, doi: 10.1038/s41746-023-00927-3.

[33] A. Alaa, B. van Breugel, E. S. Saveliev, and M. van der Schaar, "How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models," in Proc. 39th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 162, 2022, pp. 290–306.

[34] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 5, pp. 845–869, May 2014, doi: 10.1109/TNNLS.2013.2292894.

[35] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," J. Artif. Intell. Res., vol. 70, pp. 1373–1411, 2021, doi: 10.1613/jair.1.12125.

[36] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in Proc. Neural Information Processing Systems Track on Datasets and Benchmarks, vol. 1, 2021.

[37] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," Med. Image Anal., vol. 65, Art. no. 101759, Oct. 2020, doi: 10.1016/j.media.2020.101759.

[38] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, Art. no. 15, Dec. 2012, doi: 10.1145/2382577.2382579.

[39] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," Nature Commun., vol. 15, no. 1, Art. no. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

[40] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard, "Navigating the pitfalls of applying machine learning in genomics," Nature Rev. Genet., vol. 23, no. 3, pp. 169–181, Mar. 2022, doi: 10.1038/s41576-021-00434-9.

[41] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Mach. Intell., vol. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.

[42] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, Art. no. 91, Feb. 2006, doi: 10.1186/1471-2105-7-91.

[43] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," J. Clin. Epidemiol., vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[44] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," BMC Med., vol. 17, no. 1, Art. no. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[45] S. G. Finlayson et al., "The clinician and dataset shift in artificial intelligence," N. Engl. J. Med., vol. 385, no. 3, pp. 283–286, Jul. 2021, doi: 10.1056/NEJMc2104626.

[46] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," npj Digit. Med., vol. 5, no. 1, Art. no. 48, Apr. 2022, doi: 10.1038/s41746-022-00592-y.

[47] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.

[48] J. C. C. Kwong et al., "Standardized reporting of machine learning applications in urology: The STREAM-URO framework," Eur. Urol. Focus, vol. 7, no. 4, pp. 672–682, Jul. 2021, doi: 10.1016/j.euf.2021.07.004.

[49] R. F. Wolff et al., "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," Ann. Intern. Med., vol. 170, no. 1, pp. 51–58, Jan. 2019, doi: 10.7326/M18-1376.

[50] R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," BMJ, vol. 368, Art. no. m441, Mar. 2020, doi: 10.1136/bmj.m441.

[51] A. C. Davison and D. V. Hinkley, Bootstrap Methods and Their Application. Cambridge, U.K.: Cambridge University Press, 1997, doi: 10.1017/CBO9780511802843.

[52] L. Rubini, P. Soundarapandian, and P. Eswaran, "Chronic Kidney Disease," UCI Machine Learning Repository, 2015. [Online]. Available: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease. doi: 10.24432/C5G020.

[53] D. F. Andrews and A. M. Herzberg, Data: A Collection of Problems from Many Fields for the Student and Research Worker. New York, NY, USA: Springer-Verlag, 1985, doi: 10.1007/978-1-4612-5098-2.