ISSN: 2584-0355
DOI: doi.org/10.51219/MCCRJ/Inass-Bargaa/543

URF PUBLISHERS
connect with research world

Medical & Clinical Case Reports Journal

https://urfpublishers.com/journal/case-reports

Vol: 4 & Iss: 4

Research Article

Diagnostic Accuracy of Artificial Intelligence-Driven Predictive Algorithms for Differentiating Subtypes of Primary Aldosteronism

Otmane Hadji[1], Adam Hadji[2], Inass Bargaa[3]* and Eman Zuhair A. Alsalman[4]

[1]Department of Endocrinology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

[2]Xi'an Jiaotong University, Xi'an, China

[3]Department of Ophthalmology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

[4]The Second Affiliated Hospital of Xi'an Jiaotong University, China

Citation: Hadji O, Hadji A, Bargaa I, Alsalman EZA. Diagnostic Accuracy of Artificial Intelligence-Driven Predictive Algorithms for Differentiating Subtypes of Primary Aldosteronism. *Medi Clin Case Rep J* 2026;4(4):2077-2088. DOI: doi.org/10.51219/MCCRJ/Inass-Bargaa/543

Received: 13 November, 2026; **Accepted:** 24 October, 2026; **Published:** 26 November, 2026

***Corresponding author:** Inass Bargaa, Department of Ophthalmology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Copyright: © 2026 Bargaa I, et al., This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

Background: Accurate differentiation between unilateral and bilateral primary aldosteronism (PA) is essential for guiding treatment decisions, including potentially curative adrenalectomy or long-term pharmacological therapy. Although adrenal venous sampling (AVS) remains the reference standard for adrenal lateralization, its invasiveness, technical complexity, and limited availability highlight the need for alternative diagnostic approaches. This systematic review aimed to evaluate the diagnostic performance, validation strategies, methodological quality, and clinical applicability of machine learning (ML) algorithms for PA subtype classification.

Methods: A systematic literature search was conducted across six databases through October 2026 in accordance with PRISMA 2020 guidelines. Eligible studies evaluated ML-based algorithms for PA subtype classification or related diagnostic tasks against AVS, surgical histopathology, or a validated imaging-based lateralization index. Two reviewers independently performed study selection. Data extraction and risk-of-bias assessment were conducted by one reviewer and verified by a second reviewer using PROBAST and QUADAS-2. Owing to methodological and clinical heterogeneity, findings were synthesized narratively.

Results: Fifteen retrospective studies with reported sample sizes ranging from 89 to 615 were included. Among the 13 studies reporting the area under the receiver operating characteristic curve (AUC), 10 achieved at least one AUC ≥0.85, and four exceeded 0.95, predominantly in derivation or internal validation cohorts. Only four studies performed independent external validation. All four studies reporting comparable AUC values across derivation or internal and external or temporal validation cohorts demonstrated reduced discrimination, with AUC decreases ranging from 0.015 to 0.236. Proposed diagnostic algorithms suggested that AVS could potentially be avoided in 43.7%–64.8% of patients. PROBAST identified four studies at high risk of bias

and 11 at unclear risk, with none classified as low risk. Major methodological limitations included the absence of reported model calibration, limited external validation, relatively small sample sizes, and heterogeneous classification tasks.

Conclusions: Machine learning-based predictive algorithms demonstrate promising discriminatory performance in PA subtype classification. However, the retrospective nature of the available evidence, limited external validation, and methodological limitations preclude their adoption as a replacement for AVS. Their potential clinical utility lies in supporting diagnostic decision-making and facilitating patient selection for AVS. Prospective, multicenter studies incorporating independent external validation, robust calibration assessment, and standardized reporting in accordance with TRIPOD+AI are warranted before routine clinical implementation.

Keywords: Primary aldosteronism; Adrenal venous sampling; Artificial intelligence; Machine learning; Diagnostic accuracy; Adrenal lateralization; Predictive modeling; Clinical decision support; Endocrine hypertension.

Introduction

Primary aldosteronism (PA) is an important and often underdiagnosed cause of secondary hypertension, characterized by autonomous overproduction of aldosterone by the adrenal cortex[1]. Excess aldosterone causes sodium retention, potassium excretion and renin suppression and the resulting hypertension is frequently resistant to conventional drug therapy[2]. Prevalence estimates range from 5% to 15% in primary care and exceed 20% in resistant hypertension[3]. The two main subtypes are unilateral aldosterone-producing adenoma (APA) and bilateral adrenal hyperplasia (BAH) and the distinction determines treatment. Unilateral disease is usually treated with laparoscopic adrenalectomy, which can cure biochemical abnormality and improve blood pressure control[4]. Scoring systems estimate the likelihood of hypertension cure before surgery[5]. Bilateral hyperplasia is treated with mineralocorticoid receptor antagonists, a lifelong medical therapy to control blood pressure and reduce the excess cardiovascular risk of PA[6].

The reference standard for separating APA from BAH is adrenal venous sampling (AVS), an invasive procedure that requires selective catheterization of both adrenal veins[7]. The success rate of AVS can be as low as 50% to 70% in less experienced centers and complications include adrenal vein rupture and groin hematoma[8]. AVS is also costly and needs specialized interventional radiology expertise, which limits access[9]. Where AVS is unavailable, patients may be managed on the basis of cross-sectional imaging, which is unreliable for subtype classification: computed tomography (CT) and magnetic resonance imaging (MRI) can misclassify micro-adenomas or normal-appearing adrenals in up to 40% of cases, which may lead to unnecessary adrenalectomy or a missed surgical cure[10].

Artificial intelligence (AI), in particular machine learning (ML) and deep learning (DL), can identify non-linear relationships in high-dimensional data[11]. In PA, investigators have used these methods to build models that predict subtypes from clinical, biochemical and imaging data that are available without AVS[12]. The algorithms range from classical classifiers such as random forests (RF) and support vector machines (SVM) to neural networks (NN) and deep learning architectures and single-center retrospective cohorts have reported areas under the receiver operating characteristic (ROC) curve (AUC) above 0.85[13].

Several gaps limit clinical use. Few models have been validated externally and many are developed and tested in a single institution with small samples, which raises concern about generalizability across patient populations, disease severity and imaging protocols[14]. Studies also differ in the reference standard, the clinical features used, the extraction of CT radiomic features and the metrics reported (AUC, sensitivity, specificity and predictive values)[15]. Reporting guidelines such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD), recently updated for artificial intelligence models (TRIPOD+AI)[16] and risk-of-bias tools such as the Prediction model Risk of Bias Assessment Tool (PROBAST)[17] were developed because prediction-model studies are often incompletely reported and prone to bias.

Related reviews exist, but to our knowledge none is restricted to machine learning classifiers for PA subtyping or used PROBAST. Ng et al. reviewed and meta-analyzed 35 studies of tools that predict bilateral PA without AVS[18], reviewed 28 studies of algorithms that predict unilateral PA to select patients for AVS, using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) to assess risk of bias[19] and meta-analyzed algorithms and imaging that might allow surgical referral without AVS[20]. These reviews grouped tools by the type of predictor used (clinical, biochemical and imaging) rather than by modeling approach[18,19]. A broader review described artificial intelligence across the PA care pathway, from screening to treatment and noted small samples and a lack of real-world validation[21]. We therefore focused on machine learning classifiers for PA subtype and appraised them with PROBAST as well as QUADAS-2, with attention to the reference standard, the type of validation (internal versus external) and the reported performance in each study.

This systematic review evaluates the diagnostic accuracy of AI-driven predictive algorithms for differentiating PA subtypes, describes how the included studies differ in design, inputs, reference standards and validation and asks whether and under what conditions, these algorithms could complement AVS.

Methods

We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 statement[22] and completed its checklist (supplementary file). The protocol, which defined the research question, search strategy, eligibility criteria, data extraction and quality assessment, was dated 1 June 2026 and registered on 4 June 2026, after searching and screening had been completed but before data extraction and risk-of-bias assessment, in the International Prospective Register of

Systematic Reviews (PROSPERO; CRD420261415362; https://www.crd.york.ac.uk/PROSPERO/view/CRD420261415362).

Search strategy

We searched PubMed, Scopus, Web of Science, IEEE Xplore, ScienceDirect and Google Scholar; Google Scholar was used to find conference papers and journal articles not indexed in the other databases. Only published studies were sought and any preprints retrieved were excluded under exclusion criterion 7. The last search was run in October 2026. The search combined four concept blocks (population, algorithm, subtype and diagnostic outcome and accuracy measures) derived from the Population, Intervention, Comparison, Outcome (PICO) framework, with Medical Subject Headings in PubMed and title, abstract and keyword fields in Scopus and Web of Science. It had no language or publication-date limits, but we restricted inclusion to studies with an English full text or a detailed structured abstract. The full strategies are in Supplementary **(Table S2)**.

Classification of studies by input data

We grouped the included studies by the data given to the model: (i) clinical and biochemical inputs only; (ii) clinical inputs plus visual CT findings or AVS-derived values; and (iii) radiomic or deep-learning imaging features plus clinical data. A study was placed in group (ii) if any predictor was a CT finding read visually (for example nodule diameter) or was derived from AVS. Radiomic features are quantitative texture and shape descriptors extracted from CT, MRI or positron emission tomography (PET)/CT images. Studies that used imaging alone without a machine learning classifier or that applied rule-based imaging criteria, were excluded under the eligibility criteria below.

Eligibility criteria

Studies were eligible if they (1) evaluated the diagnostic accuracy of an AI or machine learning algorithm (for example neural networks, decision trees, support vector machines, random forests, gradient boosting or ensemble methods) for differentiating PA subtypes, principally unilateral disease (APA) versus bilateral disease (BAH); (2) used AVS, surgical histopathology or a validated imaging-based lateralization index as the reference standard; (3) reported sufficient quantitative data on model performance, such as sensitivity, specificity, predictive values, accuracy, F-score or AUC; and (4) were original research articles in peer-reviewed journals or full-text conference proceedings.

We excluded studies that (1) used only rule-based or non-AI statistical methods, such as simple logistic regression without machine learning optimization or threshold criteria from univariate analysis; (2) assessed imaging features without an AI algorithm performing the classification; (3) were case reports, narrative reviews, editorials, commentaries, letters or animal or in vitro studies; (4) had overlapping or duplicate patient populations when a more comprehensive or more recent report from the same cohort was available; (5) did not describe the model development or validation dataset clearly, that is, gave no training/test split, cross-validation strategy or independent validation cohort; (6) had no English full text or detailed structured abstract; or (7) were preprints (including arXiv articles) or non-peer-reviewed technical reports.

Study selection

Two reviewers independently screened titles and abstracts in the Rayyan web application, which supports blinded parallel review and conflict management. Disagreements were resolved by consensus, with a third reviewer consulted when needed. We retrieved the full texts of all records that were potentially eligible or unclear from the abstract. The same two reviewers independently assessed each full text against the eligibility criteria, resolved discrepancies by discussion or arbitration by the third reviewer and recorded the reasons for exclusion for the PRISMA flow diagram **(Figure 1)**.

Data extraction and synthesis

From each included study we extracted the design, sample size, classification task, algorithm, input data, reference standard, validation approach and reported performance metrics, using full texts. One author (O.H.) extracted the data and a second author (I.B.) double-checked the extracted data against the full texts; disagreements were resolved by discussion. The registered protocol specified independent extraction by two reviewers; this change is reported under Registration and protocol amendments. The registered primary outcomes were AUC, sensitivity, specificity, positive and negative predictive values and the data needed to construct a 2×2 contingency table; secondary outcomes were calibration (for example the Brier score), F1-score, accuracy and the proportion of patients in whom AVS could be avoided with an integrated diagnostic flowchart. As prespecified in the protocol, because populations, inputs, reference standards and metrics differed, we did not pool results and synthesized the studies narratively. We report confidence intervals (CI) and standard deviations (SD) as given in the studies.

Risk of bias assessment

We assessed risk of bias and applicability with two tools, because the included studies developed or validated prediction models and also reported diagnostic accuracy against a reference standard. PROBAST[17] was the primary tool. It has four domains (participants, predictors, outcome and analysis) and was applied to each model development or validation study. QUADAS-2[23] was used for the diagnostic-accuracy aspects of each study (patient selection, index test, reference standard and flow and timing, with applicability concerns for the first three), treating the machine learning model as the index test. Each domain was rated low, high or unclear. Overall judgment was low only if all domains were low, high if at least one domain was high and unclear otherwise; the same rule was applied to QUADAS-2 risk of bias and applicability. Ratings were based on the full texts of the included studies. One author (O.H.) rated each study and a second author (I.B.) double-checked the ratings; disagreements were resolved by discussion. This differs from the registered protocol, which specified independent assessment by two reviewers (see Registration and protocol amendments). The use of an AI assistant is described under Use of artificial intelligence.

We used the original PROBAST rather than its update for artificial intelligence models (PROBAST+AI)[24]. PROBAST applicability was not rated separately and was assessed through the QUADAS-2 applicability concerns. Ratings were used to describe the evidence narratively. No study was excluded because of its rating and we did not weight results by rating. No

funnel plot or Egger test was done because there was no meta-analysis, so risk of bias due to missing results was not formally assessed. We did not formally assess the certainty of the evidence (for example with the Grading of Recommendations, Assessment, Development and Evaluation approach) because we did not pool results. After the main assessment we ran one post hoc sensitivity analysis that removed the studies rated high risk on PROBAST (see Results and the Subsection Registration and protocol amendments below).

Registration and protocol amendments

The review was registered in PROSPERO (CRD420261415362) on 4 June 2026, after the searches and title, abstract and full-text screening had been completed but before data extraction and risk-of-bias assessment began. The full protocol, dated 1 June 2026, is available through the PROSPERO record. The analysis excluding studies at high risk of bias on PROBAST is a post hoc sensitivity analysis. It was not part of the registered protocol and is a deviation from it. The following further deviations from the registered protocol occurred. First, data extraction and risk-of-bias assessment were planned to be done independently by two reviewers; instead, one reviewer (O.H.) performed them and a second (I.B.) checked them against the full texts. Second, the protocol planned two synthesis groups (models using clinical and biochemical features only and models integrating imaging data); we divided the imaging group into visual CT findings or AVS-derived values (group ii) and radiomic or deep-learning imaging features (group iii), because these inputs differ in how they are obtained and in whether they require AVS. Third, PROBAST applicability was not rated separately and was covered by the QUADAS-2 applicability judgments. Fourth, three included studies addressed classification tasks that extend beyond the registered population of patients with confirmed PA requiring differentiation of APA from BAH: PA versus non-functioning adenoma, histopathological classes including non-functional adenoma and APA prediction within unilateral adrenal adenomas. They were retained because they applied machine learning to subtype-related classification against an eligible reference standard, but their results are not directly comparable with those of the other studies. Fifth, positive and negative predictive values and 2×2 contingency data, registered as primary outcomes, were not tabulated; Table 3 presents the performance metrics as reported in the studies. Sixth, the last search was run in October 2026, after the review end date given in the PROSPERO record (16 July 2026).

Results

Study selection

The database search identified 242 records. After removal of 116 duplicates, 126 records were screened by title and abstract and 89 were excluded. We sought the full texts of the remaining 37 reports and retrieved all of them. Twenty-two were excluded after full-text assessment: insufficient data to extract performance metrics (n = 8), non-AI statistical methods only (n = 5), no clear reference standard (n = 3), overlapping populations (n = 2), conference abstract with insufficient detail (n = 2), preprint reporting the same cohort as an included article (n = 1) and outside the scope of the review (n = 1). Fifteen studies were included (Figure 1). The preprint (Mizutani 2025) reports the same cohort as an included article, has no validation and is not peer reviewed (exclusion criteria 4, 5 and 7). The study outside the scope (Lazzarini 2015) classified APA against all other hypertensive patients, not PA subtypes, with a composite reference standard that was not AVS in all patients (eligibility criteria 1 and 2). All excluded reports are listed with reasons in Supplementary **(Table S3)**.

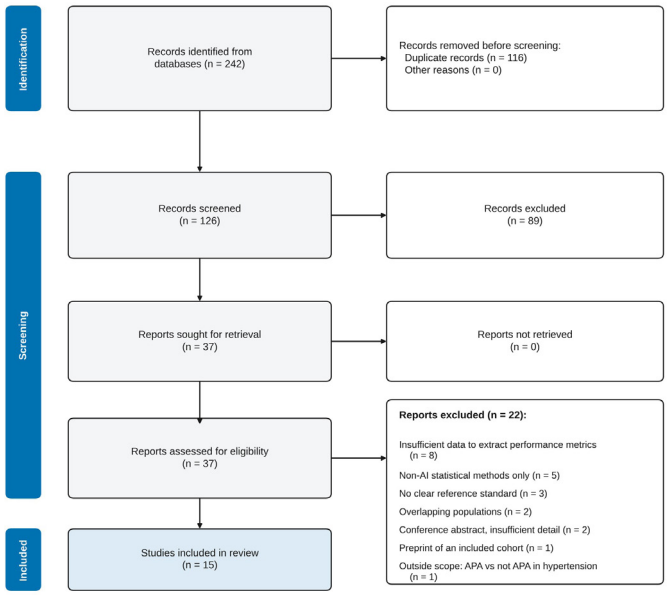

Figure 1: PRISMA 2020 flow diagram of study selection, from database searches (n = 242 records) to the 15 included studies. AI, artificial intelligence; APA, aldosterone-producing adenoma.

Publication trend

The 15 included studies were published between 2020 and 2026 (Figure 2; years are given in the study labels in Table 1). Two studies were published in 2020, two in 2021, three in 2022, two in 2023, one in 2024, one in 2025 and four in 2026 up to October. Thirteen of the 15 studies were published in 2021 or later. With so few studies per year, the counts do not support a conclusion about a trend.

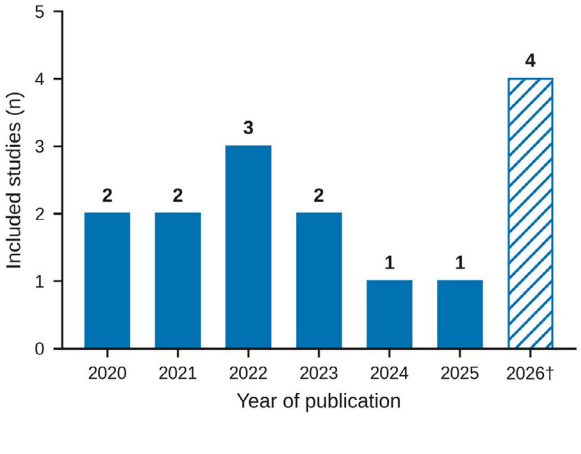

Figure 2: Included studies by year of publication (n = 15). † 2026: publications to date (October 2026); the hatched bar marks the incomplete year.

Characteristics of included studies

(Table 1) summarizes the 15 studies. Two studies by Burrello et al. were published in 2020 and are labeled 2020a (European Journal of Endocrinology)[25] and 2020b (Journal of Clinical Endocrinology and Metabolism)[26]. All studies were retrospective analyses; one drew on prospectively collected data, the German

Conn's Registry[27]. The number of participants ranged from 89 to 615. The reference standard was AVS in 13 studies and aldosterone synthase (CYP11B2) immunohistochemistry in two[28,29]. Four studies tested the model in an independent external cohort[26,29-31], one used a temporal validation cohort[32] and one used a multicenter scheme (cross-center domain adaptation[33]). Nine studies used internal validation only, with random or stratified splits, cross-validation or both[25,27,28,34-39].

Table 1: Characteristics of the included studies.

Study	Design	Total n	Train n	Test n	Ext. n	Classification task	Reference standard	Validation approach
(i) Clinical and biochemical inputs only								
Kaneko[30]	Retrospective	350[a]	183	46	121	Unilateral vs bilateral PA	AVS (with ACTH)	80/20 split, 10-fold CV, external cohort
Shi[31]	Retrospective; public data plus external	208[b]	115	50	43	UPA vs BPA	AVS	70/30 split plus external cohort
Tamaru[39]	Retrospective, single-center	127[c]	NR	NR	NA	PA subtype	AVS	80/20 split
Kitamoto[33]	Retrospective, multicenter (3 institutions)	210	NR	NR	NA	Need for AVS in PA	Segment-selective AVS	Cross-center domain adaptation
(ii) Clinical inputs plus visual CT findings or AVS-derived values								
Burrello[25]	Retrospective, single-center	158	110	48	NA	UPA vs BPA[e]	AVS	Random split, 10-fold CV
Burrello[26]	Retrospective	333	150	65	118	LPA vs BPA	AVS	Internal plus external cohort
Kaneko[36]	Retrospective; 4 centers	615	492	123	NA	UPA vs BPA	AVS (with ACTH)	80/20 holdout, 10-fold CV
Luo[37]	Retrospective, single-center	235[f]	NR	NR	NA	UPA vs BPA	AVS	70/30 split, 10-fold CV
Mizutani[38]	Retrospective, single-center	274	205	69	NA	UPA vs BPA	Successful AVS	Holdout
(iii) Radiomic or deep-learning imaging features plus clinical data								
He[35]	Retrospective	90	62	28	NA	APA prediction in unilateral adrenal adenoma	AVS	Internal validation
Mansour[27]	Retrospective; prospective registry	269	215	54	NA	Left, right or bilateral aldosterone excess	AVS	Stratified 80/20 split
Chen[34]	Retrospective	158	NA	NA	NA	Stage 1: unilateral vs bilateral; stage 2: side	AVS	Repeated stratified 5-fold CV
Zhao H[32]	Retrospective, single-center	170[g]	120	NA	50	Dominant vs non-dominant CT-negative adrenal gland	AVS	Temporal validation
Zhao Y[28]	Retrospective, single-center	89	62	27	NA	PA vs non-functioning adenoma	CYP11B2 IHC	Stratified 5-fold CV plus test set
Zhou[29]	Retrospective	190	140	NA	50	Classical vs nonclassical vs non-functional (histopathology)	CYP11B2 IHC	5-fold CV plus external cohort

[a]Modeling cohort 229 (91 unilateral, 138 bilateral) plus external cohort 121.

[b]Modeling cohort 165 (71 UPA, 94 BPA) plus 43 consecutive external patients (20 UPA, 23 BPA); AVS with ACTH in the modeling data and without ACTH in the external cohort.

[c]154 enrolled; 127 analyzed.

[d]Burrello 2020a was published in the European Journal of Endocrinology and Burrello 2020b in the Journal of Clinical Endocrinology and Metabolism.

[e]Patients with contralateral suppression at AVS.

[f]276 enrolled; 235 analyzed.

[g]Unit of analysis is the adrenal gland: 85 dominant and 85 non-dominant CT-negative glands, balanced by design.

Abbreviations: Train, training or modeling cohort; Test, internal test or validation cohort; Ext., external or temporal cohort; ACTH, adrenocorticotropic hormone; APA, aldosterone-producing adenoma; AVS, adrenal venous sampling; BPA, bilateral PA; CT, computed tomography; CV, cross-validation; CYP11B2, aldosterone synthase; IHC, immunohistochemistry; LPA, lateralized PA; NA, not applicable; NR, not reported; PA, primary aldosteronism; UPA, unilateral PA.

Models and performance by input type

(Tables 2 and 3) lists the algorithms, inputs and main predictors and Table 3 the reported performance, grouped by input type. Random forest was named in eight studies[25-27,30,31,33,38,39] and neural networks, deep learning or a multilayer perceptron in three[33,34,38]. Six studies addressed tasks other than separating unilateral from bilateral disease: APA prediction within unilateral adenomas[35], identification of patients who need AVS[33], lateralization into left, right and bilateral aldosterone excess[27], identification of the dominant gland among CT-negative adrenal glands[32], PA versus non-functioning adenoma[28] and classification of histopathology, using CYP11B2 immunohistochemistry as the reference[29]. The other nine classified PA subtypes, mainly unilateral versus bilateral disease. Thirteen studies reported an AUC **(Figure 3)**, with values from 0.56 to 0.990; two reported none[33,39]. Detailed descriptions of each study are in Supplementary Text S1.

Four studies used clinical and biochemical inputs only[30,31,33,39]. Random Forest reached an AUC of 0.990 in the test set of Kaneko et al. (21 clinical variables)[30] and 0.938 in Shi et al. (10 variables)[31]. A three-variable random forest had an AUC of 0.950 in the test set but 0.826 (95% CI 0.723–0.930) in an independent external cohort[30] and the AUC in Shi et al. fell to 0.887 externally[31]. A flowchart from Kitamoto et al. would have required AVS in 35.2% of patients, with an accuracy of 94.1%[33] and Tamaru, et al. reported no extractable performance metrics[39].

Table 2: Algorithms, input data and main predictors.

Study	Algorithm(s)	Input data	Main predictors reported
(i) Clinical and biochemical inputs only			
Kaneko[30]	LR, SVM, RF, GBDT	21 clinical variables	K, PAC, Na (feature-selected RF)
Shi[31]	RF	10 clinical and biochemical variables, including CCT and SIT results	PAC after SIT, ARR, ARR after CCT
Tamaru[39]	RF	Clinical and loading-test results	NR
Kitamoto[33]	LR, RF, DL	36 clinical markers (missing data imputed)	NR
(ii) Clinical inputs plus visual CT findings or AVS-derived values			
Burrello[25]	LDA, RF; 19-point score	Clinical, biochemical, CT findings, AVS	Aldosterone (screening, after test), lowest K, ipsilateral and contralateral CT findings, contralateral ratio at AVS
Burrello[26]	LDA, RF; 20-point score	Clinical, biochemical, CT findings	Aldosterone (screening, after test), lowest K, nodule presence and diameter, CT result
Kaneko[36]	LightGBM	Clinical, biochemical, CT nodule diameter	PAC (baseline, after SIT, after CCT), lowest K, nodule diameter
Luo[37]	XGBoost, LASSO (sex-specific)	Clinical, biochemical, CT-derived	Men: VFA, K, SBP, PAC. Women: K, lipids, PAC after SIT, CT nodules
Mizutani[38]	Five ML algorithms (LR, SVM, RF, GBDT, MLP)	196 features	CCT90-PAC, CT tumor size (top RF features)
(iii) Radiomic or deep-learning imaging features plus clinical data			
He[35]	LASSO and multivariable LR (nomogram)	Unenhanced CT radiomics, clinical	Radscore, age, sex, K, ARR
Mansour[27]	RF	Native CT radiomics, with and without clinical parameters	Lowest baseline K, SIT result (clinical parameters)
Chen[34]	Neural networks	Triple-phase CT radiomics, clinicobiochemical	NR
Zhao H[32]	Six conventional ML models (unspecified)	CT radiomics (107 features, 10 selected), clinical	K, aldosterone excretion, uric acid, IVSd
Zhao Y[28]	SVM	Clinical, ^{18}F-pentixafor PET/CT radiomics, deep-learning features	Lesion-to-adrenal ratio, SUVmax (SHAP)
Zhou[29]	ElasticNet, SHAP	CT radiomics, clinical	Lowest K, Na, radiomic scores

[a]Visual CT finding (nodule presence, size, laterality), not radiomics. The model of Burrello[25] also uses the AVS contralateral ratio, so it requires AVS data.
[b]Placed in group (ii) because a visually read CT measurement is a predictor.
[c]Uses CT-derived visceral fat area and nodule characteristics, not radiomics.
Abbreviations: ARR, aldosterone-to-renin ratio; AVS, adrenal venous sampling; CCT, captopril challenge test; CT, computed tomography; DL, deep learning; GBDT, gradient-boosted decision trees; IVSd, interventricular septal diameter in diastole; K, serum potassium; LASSO, least absolute shrinkage and selection operator; LDA, linear discriminant analysis; LR, logistic regression; ML, machine learning; MLP, multilayer perceptron; Na, serum sodium; NR, not reported; PAC, plasma aldosterone concentration; PET, positron emission tomography; RF, random forest; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; SIT, saline infusion test; SUVmax, maximum standardized uptake value; SVM, support vector machine; VFA, visceral fat area.

Table 3: Reported diagnostic performance.

Study	Cohort or model	AUC (95% CI or SD)	Accuracy, %	Sensitivity, %	Specificity, %	F1	AVS avoided, %
(i) Clinical and biochemical inputs only							
Kaneko[30]	Test, RF, 21 variables	0.990	95.7	NR	NR	NR	NR
	Test, RF, 3 variables	0.950	89.1	NR	NR	NR	NR
	External, RF, 3 variables	0.826 (0.723–0.930)	85.1	72.4	89.1	NR	NR
Shi[31]	Internal test	0.938	90.0	81.8	96.4	0.878	NR
	External	0.887	81.4	90.0	73.9	0.818	51.2[a]
Tamaru[39]	NR	NR	NR	NR	NR	NR	NR
Kitamoto[33b]	Flowchart	NR	94.1	NR	NR	NR	64.8
(ii) Clinical inputs plus visual CT findings or AVS-derived values							
Burrello[25c]	Combined training and validation	0.971 (0.947–0.994)	NR	95.2	96.9	NR	NR
	Validation, score	NR	93.7	94.7	90.0	NR	NR
	Flowchart	NR	98.1	NR	NR	NR	77.2[d]
Burrello[26e]	Score, cohort not reported	0.896	NR	91.7	79.3	NR	NR
	Flowchart	NR	96.3[f]	NR	NR	NR	43.7
	External cohort[g]	NR	NR	NR	NR	NR	NR
Kaneko[36h]	Test, SIT plus CCT	0.850 (0.740–0.959)	91.1	NR	NR	NR	NR
	Test, SIT alone	0.813 (0.682–0.944)	90.2	NR	NR	NR	NR
	Test, CCT alone	0.786 (0.648–0.924)	90.2	NR	NR	NR	NR
Luo[37]	Men, 10-fold CV	0.84 (SD 0.05)[i]	NR	NR	NR	NR	NR
	Women, 10-fold CV	0.75 (SD 0.02)[i]	NR	NR	NR	NR	NR
	All patients, 10-fold CV	0.77 (SD 0.02)[i]	NR	NR	NR	NR	NR
Mizutani[38j]	Test, RF	0.972	91.3	76.5	96.2	NR	NR
(iii) Radiomic or deep-learning imaging features plus clinical data							
He[35]	Training	0.900 (0.807–0.993)	NR	NR	NR	NR	NR
	Internal validation	0.912 (0.761–1.000)	NR	NR	NR	NR	NR
Mansour[27]	Test, radiomics alone	0.56 (0.45–0.65)[i]	NR	NR	NR	NR	NR

Study	Cohort or model	AUC (95% CI or SD)	Accuracy, %	Sensitivity, %	Specificity, %	F1	AVS avoided, %
	Test, radiomics plus clinical	0.67 (0.57–0.77)[i]	NR	NR	NR	NR	NR
Chen[34k]	Stage 1, integrated	0.778 (SD 0.050)	80.6 (SD 3.0)	NR	NR	0.748 (SD 0.052)	NR
	Stage 2, radiomics	0.831 (SD 0.087)	88.0 (SD 4.9)	NR	NR	0.819 (SD 0.062)	NR
	Overall	0.771 (SD 0.046)	77.5 (SD 3.9)	NR	NR	0.705 (SD 0.071)	NR
Zhao H[32]	Derivation	0.868 (0.804–0.932)	78.3	70.0	86.7	NR	NR
	Temporal validation	0.853 (0.750–0.956)	76.0	72.0	80.0	NR	NR
Zhao Y[28]	Test	0.907	NR	100.0	NR	0.923	NR
Zhou[29]	Training, classical	0.912	NR	NR	NR	NR	NR
	Training, nonclassical	0.923	NR	NR	NR	NR	NR
	Training, NFA	0.958	NR	NR	NR	NR	NR
	External, classical	0.807	NR	NR	NR	NR	NR
	External, nonclassical	0.735	NR	NR	NR	NR	NR
	External, NFA	0.722	NR	NR	NR	NR	NR

[a]External cohort, at a probability threshold chosen after the fact.

[b]F-scores (surgical, medical indication): logistic regression 0.89, 0.86; random forest 0.84, 0.84; deep learning 0.82, 0.84.

[c]Machine learning model accuracy 82.9% (validation) to 95.7% (training).

[d]Repeat AVS avoided.

[e]Machine learning model accuracy 79.1% to 93.0% at training and internal validation; AUC and sensitivity/specificity refer to the 20-point score.

[f]Patients correctly sent to surgery.

[g]Described as similar; not quantified.

[h]No significant difference between the models (DeLong test). Baseline characteristics alone: AUC 0.872 (0.773–0.972), accuracy 91.1%.

[i]Reported to two decimals in the source.

[j]Other algorithms had AUCs of 0.920 to 0.932.

[k]Mean (SD) from repeated stratified 5-fold cross-validation; F1 converted from percent to proportion.

Abbreviations: AUC, area under the receiver operating characteristic curve; AVS, adrenal venous sampling; CCT, captopril challenge test; CI, confidence interval; CV, cross-validation; F1, F-score; NFA, non-functional adenoma; NR, not reported; RF, random forest; SD, standard deviation; SIT, saline infusion test. Values in parentheses after an AUC are 95% confidence intervals or standard deviations, as labeled.

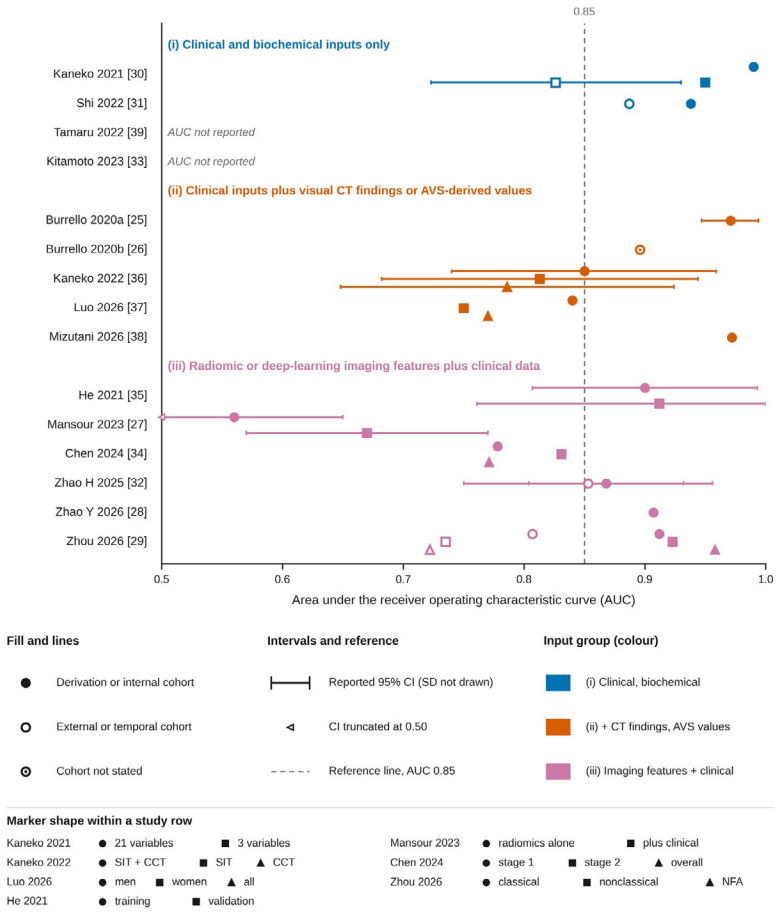

Figure 3: Area under the receiver operating characteristic curve (AUC) by study and input group (15 studies; 13 reported an AUC). Colour indicates the input group: (i) clinical and biochemical inputs only (blue), (ii) clinical inputs plus visual CT findings

or AVS-derived values (orange) and (iii) radiomic or deep-learning imaging features plus clinical data (pink). Filled markers show derivation or internal cohorts (training, test or cross-validation; for Burrello[25], the combined training and validation cohort), open markers show external or temporal cohorts and the marker with a central dot shows an AUC for which the cohort was not stated[26]. Whiskers show reported 95% CIs; ± SD values are not drawn. The lower limit of the Mansour 2023 radiomics-alone CI (0.45) is truncated at 0.50 (left-pointing triangle). The dashed vertical line marks AUC 0.85. Marker shapes (circle, square, triangle) distinguish AUC values within a study, as listed in the key; colour is never the only cue. AVS, adrenal venous sampling; CCT, captopril challenge test; CI, confidence interval; CT, computed tomography; NFA, non-functional adenoma; SD, standard deviation; SIT, saline infusion test.

Five studies added visual CT findings or AVS-derived values to clinical inputs[25,26,36-38]. The two Burrello studies-built point scores from CT findings (nodule presence, size and laterality) and biochemical values. The score of Burrello 2020b had an AUC of 0.896 and its flowchart would have avoided AVS in 43.7% of patients; external performance was described as similar but was not quantified[26]. The score of Burrello 2020a had an AUC of 0.971 (95% CI 0.947–0.994) in the combined training and validation cohort, that is, apparent performance. It also uses the contralateral ratio from AVS, so it needs AVS data and helps to interpret incomplete AVS but does not replace it[25]. Kaneko, et al. found that a model using both the saline infusion test and the captopril challenge test (AUC 0.850) did not differ significantly from models using either test alone (0.813 and 0.786)[36]. Mizutani et al. reported an AUC of 0.972 for a random forest on 196 endocrine features in a test set of 69 patients[38]. Challenge-test results ranked among the strongest predictors in that study and in Shi, et al.[31,38]. Luo, et al. fitted sex-specific models (AUC 0.84 in men and 0.75 in women, compared with 0.77 for a pooled model)[37].

Six studies combined radiomic or deep-learning imaging features with clinical data[27-29,32,34,35]. AUCs in the derivation or internal cohorts ranged from 0.56 to 0.958. In the German Conn's Registry study, radiomics alone gave an AUC of 0.56 (95% CI 0.45–0.65) in a test set of 54 patients and adding clinical parameters raised it to 0.67 (0.57–0.77); class-specific AUCs ranged from 0.50 to 0.61 with radiomics alone and from 0.61 to 0.73 with clinical data added[27]. Chen, et al. reported AUCs of 0.778 (SD 0.050) for the integrated model at stage 1 (unilateral versus bilateral) and 0.831 (SD 0.087) for the radiomics model at stage 2 (side)[34]. He, et al. reported 0.900 in training and 0.912 in internal validation (95% CI 0.761–1.000)[35]. Zhao H, et al. reported 0.868 (95% CI 0.804-0.932) in the derivation cohort and 0.853 (0.750-0.956) in a temporal validation cohort, for identifying the dominant gland among CT-negative adrenals in a sample balanced by design[32]. Zhao Y, et al. reported an AUC of 0.907 for PA versus non-functioning adenoma and the combined model outperformed single-source models ($P < 0.01$)[28]. Zhou, et al. reported AUCs of 0.912, 0.923 and 0.958 for three histopathology classes in the training cohort[29].

Internal and external validation

Four studies tested a model in an independent external cohort[26,29-31] and one used a temporal validation cohort[32]. Four reported the same AUC in both settings and in all four the AUC was lower outside the derivation cohort **(Figure 4)**: by 0.124 in Kaneko, et al.[30] (0.950 to 0.826), by 0.051 in Shi, et al.[31] (0.938 to 0.887), by 0.015 in the temporal cohort of Zhao H, et al.[32] (0.868 to 0.853, a slight decrease) and by 0.105 to 0.236 across the three histology classes in Zhou, et al.[29]. Accuracy fell from 89.1% to 85.1% in Kaneko et al. and from 90.0% to 81.4% in Shi, et al. Burrello, et al.[26] described external performance as similar to internal performance but gave no values. Of the six studies that used radiomic or deep-learning imaging features, only two tested the model outside the derivation cohort[29,32].

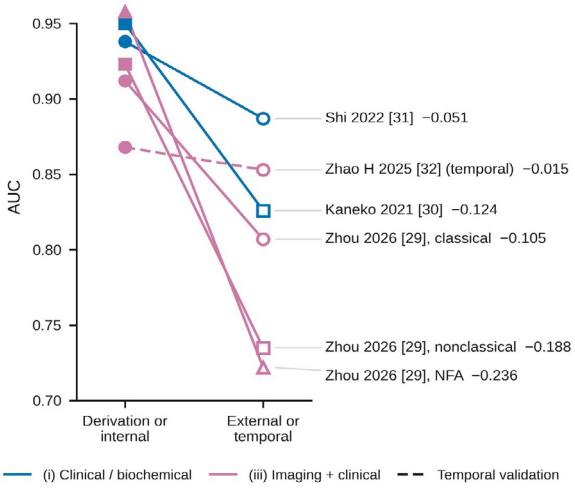

Figure 4: Change in AUC between the derivation or internal cohort (filled markers) and an external or temporal cohort (open markers) in the four studies that reported the same AUC in both settings (six lines: Shi 2022, Kaneko 2021, Zhao H 2025 and the three histology classes of Zhou 2026). The dashed line denotes temporal validation. Labels give the AUC in the external or temporal cohort minus the AUC in the derivation cohort. Burrello[26] described external performance as similar without giving a value and is not plotted. Colours follow **(Figure 3)** (blue, input group (i); pink, input group (iii)) and marker shapes identify subgroups as in Figure 3. AUC, area under the receiver operating characteristic curve; NFA, non-functional adenoma.

Risk of bias

Risk of bias was assessed for all 15 studies (**Table 4 and Figure 5**; domain-level ratings and rationales in Supplementary Table S1). No study was rated at low risk of bias overall on PROBAST: 4 studies (27%) were rated high and 11 (73%) unclear. The predictors domain was rated low in all studies and the outcome domain in 14 (unclear in Shi 2022). The participant's domain was rated high in 3 studies (Shi 2022, Zhao Y 2026 and Zhou 2026) and the analysis domain in 2; the other 13 analysis ratings were unclear. Calibration was not reported in any of the 15 studies and 11 studies were validated only internally, temporally or across centers.

The four studies at high risk were Shi 2022 (participants and analysis), Luo 2026 (analysis only) and Zhao Y 2026 and Zhou 2026 (participants, because only patients who had adrenalectomy were included). On QUADAS-2, risk of bias was low in 11 studies and high in 4 (the same four studies as on PROBAST). Applicability concerns were low in 10 studies, unclear in 1 (Shi 2022) and high in 4: Burrello 2020a needs AVS-derived data and He 2021, Zhao Y 2026 and Zhou 2026 used narrow populations.

We found no consistent link between risk of bias and performance. Of the 10 studies with at least one AUC of 0.85 or higher, 3 were rated high and 7 unclear and the three highest AUCs (0.990, 0.972 and 0.971) came from studies rated unclear. Of the 4 studies with an independent external cohort, 2 were rated high and 2 unclear. No funnel plot or Egger test was done, so risk of bias due to missing results was not assessed.

In a post hoc sensitivity analysis, after excluding the 4 studies at high risk on PROBAST from the 15, 11 studies remained; 9 reported an AUC and 7 of these had at least one value of 0.85 or higher, 2 had an independent external cohort and 1 a temporal cohort. QUADAS-2 and applicability count for this analysis are in Supplementary Text S1.

Table 4: Risk of bias and applicability of the 15 included studies on PROBAST and QUADAS-2.

	PROBAST					QUADAS-2 risk of bias					Appl.	
Study	**Part.**	**Pred.**	**Outc.**	**Anal.**	**Ov.**	**PS**	**IT**	**RS**	**FT**	**Ov.**	**Ov.**	**Main reason(s) for the ratings**
Kaneko 2021 [30]	L	L	L	U	U	L	L	L	L	L	L	No calibration reported; hyperparameters tuned by grid search; internal (10-fold CV) and external validation done.
Shi 2022 [31]	H	L	U	H	H	H	H	U	H	H	U	Public database, enrolment unclear; small external cohort (43); overfitting risk, no calibration, AUC fell externally; AVS protocols differed.
Tamaru 2022 [39]	L	L	L	U	U	L	L	L	L	L	L	Internal validation only; no calibration; small sample.
Kitamoto 2023 [33]	L	L	L	U	U	L	L	L	L	L	L	Cross-center validation, no independent external cohort; no calibration; complex imputation.
Burrello 2020a [25]	L	L	L	U	U	L	L	L	L	L	H	Internal validation only; no calibration. Needs AVS-derived contralateral ratio; patients with contralateral suppression.
Burrello 2020b [26]	L	L	L	U	U	L	L	L	L	L	L	Internal and external validation; overfitting effect calculated; no calibration.
Kaneko 2022 [36]	L	L	L	U	U	L	L	L	L	L	L	Internal validation only; no calibration; multiple comparisons without correction.
Luo 2026 [37]	L	L	L	H	H	L	H	L	H	H	L	Internal validation only (70/30, 10-fold CV), no clearly independent test result; small sex subgroups; no calibration; 41 of 276 excluded unexplained.
Mizutani 2026 [38]	L	L	L	U	U	L	L	L	L	L	L	Holdout validation only; no calibration; 196 features for 205 training cases.
He 2021 [35]	L	L	L	U	U	L	L	L	L	L	H	Internal validation only; small sample; no calibration. Only patients with a unilateral adrenal adenoma.
Mansour 2023 [27]	L	L	L	U	U	L	L	L	L	L	L	Internal validation only; no calibration; no hyperparameter tuning.
Chen 2024 [34]	L	L	L	U	U	L	L	L	L	L	L	Internal validation only; no calibration; complex model with small sample.
Zhao H 2025 [32]	L	L	L	U	U	L	L	L	L	L	L	Temporal validation only, no external; no calibration; complex iterative feature selection.
Zhao Y 2026 [28]	H	L	L	U	H	H	L	L	L	H	H	Surgical patients only, no bilateral disease. Internal 5-fold CV only; small sample; no calibration.
Zhou 2026 [29]	H	L	L	U	H	H	L	L	L	H	H	Adrenalectomy patients only, no bilateral PA. Internal 5-fold CV and small external cohort; no calibration.

Abbreviations: L, low; U, unclear; H, high; Ov., overall; Part., participants; Pred., predictors; Outc., outcome; Anal., analysis; PS, patient selection; IT, index test; RS, reference standard; FT, flow and timing; Appl., applicability (overall QUADAS-2 judgment); PROBAST, Prediction model Risk Of Bias ASsessment Tool; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies 2. Overall: low only if all domains are low, high if any is high, unclear otherwise. Domain-level applicability ratings and full rationales: Supplementary Table S1.



Figure 5: Risk of bias and applicability of the 15 included studies. Rows are studies in the order of Table 1, grouped by input

type (i) to (iii) as in Figure 3. Columns show PROBAST domains (Part., participants; Pred., predictors; Outc., outcome; Anal., analysis), QUADAS-2 risk-of-bias domains (PS, patient selection; IT, index test; RS, reference standard; FT, flow and timing) and QUADAS-2 applicability concerns (PS, IT, RS). Larger circles are overall judgments. Green circle with L, low risk; amber circle with U, unclear risk; coral circle with H, high risk. The bottom row gives the number of studies rated low/unclear/high in each column. PROBAST, Prediction model Risk Of Bias ASsessment Tool; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies 2. Full rationales are in Supplementary Table S1.

Discussion

We identified 15 studies of machine learning models for classifying PA subtypes. Discrimination can be high: 10 of the 13 studies that reported an AUC had at least one value of 0.85 or higher, with the highest values for random forests on clinical or endocrine variables[30,38] and for a score that includes AVS-derived data[25]. The range was wide, however, with AUCs of 0.56 and 0.67 in the registry study[27] and 0.771-0.831 in a CT radiomics study[34], against 0.990 for a random forest on 21 clinical variables[30]. Most estimates come from derivation or internal cohorts and in the four studies that reported the same AUC in a second cohort performance fell by 0.015 to 0.236 **(Figure 4)**, so published estimates may overstate performance in new populations.

No algorithm was consistently best. Random forest, the most frequently named algorithm (8 of 15 studies), had the highest accuracy in one four-classifier comparison[30] but was outperformed by logistic regression in another[33] and performed poorly in the registry study[27]. Challenge-test results were among the strongest predictors in two studies[31,38], but in another study baseline potassium, aldosterone and sodium were most important and a model using two confirmatory tests did not differ significantly from models using one[30,36]. The benefit of sex-specific modeling was not consistent[37]. Imaging features added inconsistent value: three studies found that combining clinical data with imaging features outperformed single-source models[27,28,34], yet the highest AUCs came from models without radiomics[25,30,38]. The six imaging studies addressed different tasks and used different reference standards, so they cannot be compared directly. The Burrello models use visual CT findings, not radiomics and the 2020a model needs the contralateral ratio from AVS, so it helps to interpret incomplete AVS and does not replace it[25].

Four studies estimated how many AVS procedures a flowchart or threshold could avoid: 43.7% in Burrello[26], 51.2% in Shi et al. at a threshold chosen after the fact[31], 64.8% in the multicenter study of Kitamoto, et al.[33] and 77.2% of repeat AVS procedures in patients with contralateral suppression at AVS in Burrello[25]. All used retrospective data and none has been tested prospectively. If validated, such flowcharts could limit AVS to patients whose prediction is indeterminate or discordant, which may matter most where AVS is unavailable or success rates are low. The included studies did not test that use.

Several features of the evidence limit these conclusions. External validation was rare: 4 of 15 studies used an independent external cohort[26,29-31] and one used a temporal cohort[32]. Nine studies used internal validation only[25,27,28,34-39], which cannot detect center-specific differences in patients, imaging protocols or referral patterns and the number of participants with the target condition was small in at least two studies, for example 17 to 20 with unilateral disease in the test sets of Kaneko et al. and Mizutani, et al.[36,38]. Reference standards also differed: AVS in 13 studies and CYP11B2 immunohistochemistry in two, with AVS protocols and lateralization criteria that were not standardized across or within studies[30,31,36]. Immunohistochemistry cannot be applied to patients with bilateral disease who are treated medically[28,29]. Two studies reported no AUC[33,39] and calibration was not reported in any of the 15 studies.

No study was at low risk of bias overall on PROBAST and 4 were at high risk on QUADAS-2. Most concerns arose in the analysis domain, so the performance estimates above may be optimistic, consistent with the lower AUC seen in external and temporal cohorts. Four studies had high applicability concerns, because the model needs AVS-derived data[25] or was developed in a narrow population[28,29,35]. Their results therefore do not directly answer whether a model can replace AVS in patients with confirmed PA. Publication bias may also have inflated apparent performance.

The review has limitations. We included only studies with an English full text or a detailed structured abstract. The protocol was registered in PROSPERO after the searches and screening had been completed, although before data extraction and risk-of-bias assessment, so the search and eligibility decisions were not fully protected from knowledge of the retrieved records. Contrary to the registered protocol, data extraction and risk-of-bias assessment were done by one reviewer and checked by a second rather than independently by two reviewers, which may have increased the risk of errors. Three included studies addressed classification tasks beyond the registered population[28,29,35], which limits the comparability of their results. The post hoc sensitivity analysis removing studies at high risk of bias on PROBAST is a deviation from the registered protocol (see Registration and protocol amendments). Our search may have missed studies that used other terms. QUADAS-2 was designed for conventional diagnostic tests and was adapted here to machine learning, we used the original PROBAST rather than PROBAST+AI[24] and the split between high and unclear ratings in the analysis domain (13 studies unclear) depends on how strictly the tool is applied. We did not grade the certainty of the evidence formally and heterogeneity led us to a narrative synthesis without pooled estimates.

Future studies should be prospective and multicenter, follow standardized protocols for data collection, model development and reporting (TRIPOD+AI[16]) and use pre-registered analysis plans to limit overfitting. They should compare models directly with conventional clinical decision rules and with AVS, ideally in a controlled design and state in advance the intended role of a model: a screening tool to reduce unnecessary AVS, a replacement for AVS in selected patients or a triage tool that ranks patients for AVS. The added value of functional imaging, such as PET/CT with tracers targeting the C-X-C chemokine receptor type 4 (CXCR4), rests on one small study[28] and needs testing in larger cohorts. Other open questions are performance in minority populations and in atypical presentations such as unilateral hyperplasia or aldosterone-producing carcinoma and the temporal stability of models.

Conclusion

In 15 retrospective studies, machine learning models separated PA subtypes with an AUC of 0.85 or higher in 10 of the 13 studies that reported an AUC, mostly in derivation or internal cohorts. Only 4 studies tested a model in an independent external cohort and the AUC was lower outside the derivation cohort in all 4 studies that reported it in both settings. Flowcharts or thresholds based on three models were estimated to avoid AVS in 43.7% to 64.8% of patients in retrospective data. No study was at low risk of bias on PROBAST, mainly because of the analysis domain. Challenge-test results were among the strongest predictors in two studies, whereas the benefit of adding imaging features was inconsistent. Because no study was rated at low risk of bias overall on PROBAST, confidence in this evidence is limited, although certainty was not formally graded and it does not support replacing AVS with machine learning predictions in routine practice. Prospective multicenter studies with pre-registered protocols, consistent reference standards and reporting according to TRIPOD+AI are needed before clinical use.

References

1. Brown JM, Siddiqui M, Calhoun DA, et al. The unrecognized prevalence of primary aldosteronism: a cross-sectional study. Ann Intern Med 2020;173(1):10-20.
2. te Riet L, van Esch JHM, Roks AJM, et al. Hypertension: renin-angiotensin-aldosterone system alterations. Circ Res 2015;116(6):960-975.
3. Douma S, Petidis K, Doumas M, et al. Prevalence of primary hyperaldosteronism in resistant hypertension: a retrospective observational study. Lancet 2008;371(9628):1921-1926.
4. Steichen O, Zinzindohoué F, Plouin PF, et al. Outcomes of adrenalectomy in patients with unilateral primary aldosteronism: a review. Horm Metab Res 2012;44(3):221-227.
5. Marzano L, Husain-Syed F, Reis T, et al. Assessment of performance of stratum-specific likelihood ratios of the aldosteronoma resolution score for predicting hypertension cure after adrenalectomy for primary aldosteronism: a systematic review and meta-analysis. J Hum Hypertens 2023;37:532-541.
6. Chevalier B, Vantyghem MC, Espiard S. Bilateral adrenal hyperplasia: pathogenesis and treatment. Biomedicines 2021;9(10):1397.
7. Blondin D, Quack I, Haase M, et al. Indication and technical aspects of adrenal blood sampling. Rofo 2015;187(1):19-28.
8. Daunt N. Adrenal vein sampling: how to make it quick, easy and successful. Radiographics 2005;25(1):143-158.
9. Turcu AF, Auchus RJ. Approach to the patient with primary aldosteronism: utility and limitations of adrenal vein sampling. J Clin Endocrinol Metab 2021;106(4):1195-208.
10. Zhou Y, Wang D, Jiang L, et al. Diagnostic accuracy of adrenal imaging for subtype diagnosis in primary aldosteronism: systematic review and meta-analysis. BMJ Open 2020;10(12):e038489.
11. Badawy M, Ramadan N, Hefny HA. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. J Electr Syst Inf Technol 2023;10:40.
12. Yu D, Zhang J, Li X, et al. Developing the novel diagnostic model and potential drugs by integrating bioinformatics and machine learning for aldosterone-producing adenomas. Front Mol Biosci 2024;10:1308754.
13. Eisenhofer G, Durán C, Cannistraci CV, et al. Use of steroid profiling combined with machine learning for identification and subtype classification in primary aldosteronism. JAMA Netw Open 2020;3(9):e2016209.
14. de Freminville JB, Amar L, Feydy J, et al. Predicting primary aldosteronism: risk stratification-based and machine learning-based models. Eur J Prev Cardiol 2026:197.
15. Williams TA, Constantinescu G, et al. Biomarkers for subtype stratification in primary aldosteronism: current and future perspectives. Eur J Endocrinol 2025;193(5):51-64.
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.
17. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58.
18. Ng E, Gwini SM, Zheng W, Fuller PJ, Yang J. Predicting bilateral subtypes of primary aldosteronism without adrenal vein sampling: a systematic review and meta-analysis. J Clin Endocrinol Metab 2024;109(2):e837-55.
19. Ng E, Gwini SM, Zheng W, et al. Tools to predict unilateral primary aldosteronism and optimise patient selection for adrenal vein sampling: a systematic review. Clin Endocrinol (Oxf) 2025;103(1):3-12.
20. Ng E, Gwini SM, Zheng W, et al. Identifying unilateral aldosterone-producing adenomas using published algorithms and imaging: a systematic review and meta-analysis. Endocr Connect 2025;14(11):e250339.
21. Xu Y, Liu B, Huang X, et al. Artificial intelligence in primary aldosteronism: current achievements and future challenges. Front Mol Biosci 2025;12:1660588.
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
23. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-536.
24. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025;388:e082505.
25. Burrello J, Burrello A, Pieroni J, et al. Prediction of hyperaldosteronism subtypes when adrenal vein sampling is unilaterally successful. Eur J Endocrinol 2020;183(6):657-667.
26. Burrello J, Burrello A, Pieroni J, et al. Development and validation of prediction models for subtype diagnosis of patients with primary aldosteronism. J Clin Endocrinol Metab 2020;105(10):3706-3717.
27. Mansour N, Mittermeier A, Walter R, et al. Integration of clinical parameters and CT-based radiomics improves machine learning assisted subtyping of primary hyperaldosteronism. Front Endocrinol (Lausanne) 2023;14:1244342.
28. Zhao Y, Chen Y, Duan F, et al. Non-invasive prediction of CYP11B2-defined subtypes in primary aldosteronism using 18F-pentixafor PET/CT and machine learning. Acad Radiol 2026;33(5):2246-2256.
29. Zhou S, Liu Z, Fang W, Xu X, Gong X. An interpretable machine learning model for preoperative prediction of aldosterone secretion and CYP11B2 status of adrenal gland. BMC Med Imaging 2026;26:146.
30. Kaneko H, Umakoshi H, Ogata M, et al. Machine learning-based models for prediction of subtype diagnosis of primary aldosteronism using blood test. Sci Rep 2021;11:9140.
31. Shi S, Tian Y, Ren Y, Li Q, et al. A new machine learning-based prediction model for subtype diagnosis in primary aldosteronism. Front Endocrinol (Lausanne) 2022;13:1005934.
32. Zhao H, Hu P, Mao M, Li X, Wang L, Chang J. Development and validation of prediction models for special subtype of primary aldosteronism: patients with negative adrenal CT imaging. Front Endocrinol (Lausanne) 2025;16:1563748.

33. Kitamoto T, Idé T, Tezuka Y, et al. Identifying primary aldosteronism patients who require adrenal venous sampling: a multi-center study. Sci Rep 2023;13:21722.

34. Chen PT, Li PY, Liu KL, et al. Machine learning model with computed tomography radiomics and clinic biochemical characteristics predict the subtypes of patients with primary aldosteronism. Acad Radiol 2024;31:1818-27.

35. He K, Zhang ZT, Wang ZH, et al. A clinical-radiomic nomogram based on unenhanced computed tomography for predicting the risk of aldosterone-producing adenoma. Front Oncol 2021;11:634879.

36. Kaneko H, Umakoshi H, Fukumoto T, et al. Do multiple types of confirmatory tests improve performance in predicting subtypes of primary aldosteronism? Clin Endocrinol (Oxf) 2023;98:473-480.

37. Luo Y, He X, Wang F, et al. The diagnostic role of machine learning models related to visceral fat with sex-specific distribution in subtypes of primary aldosteronism. Int J Med Sci 2026;23(6):1982-1991.

38. Mizutani Y, Miyashita K, Nakamura T, et al. Machine learning-based prediction of primary aldosteronism subtype using comprehensive clinical features. Sci Rep 2026;16:10071.

39. Tamaru S, Suwanai H, Abe H, et al. Machine learning approach to predict subtypes of primary aldosteronism is helpful to estimate indication of adrenal vein sampling. High Blood Press Cardiovasc Prev 2022;29:375-83.