British Journal of Hospital Medicine

Br. J. Hosp. Med. (Lond) **2026**; 87(8): 57289
https://doi.org/10.31083/BJHM57289

Article

Comparison of the Performances of Generalized Linear Model, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine Models From Machine Learning–Based Classification Algorithms in the Diagnosis of Cushing's Syndrome

Mustafa Çakır[1], Mustafa Aydemir[2,*], Okan Oral[3], Mesut Yılmaz[4]

[1]Iskenderun Vocational School of Higher Education, Iskenderun Technical University, 31200 Iskenderun, Turkey
[2]Division of Endocrinology and Metabolism, Department of Internal Medicine, Akdeniz University Faculty of Medicine, 07058 Antalya, Turkey
[3]Department of Machine Engineering, Faculty of Engineering, Akdeniz University, 07058 Antalya, Turkey
[4]Department of Aquaculture, Faculty of Fisheries, Akdeniz University, 07058 Antalya, Turkey
*Correspondence: mustafactf@gmail.com (Mustafa Aydemir)
Academic Editor: John Alcolado
Submitted: 22 November 2025 Revised: 29 June 2026 Accepted: 15 July 2026 Published: 21 August 2026

Abstract

Aims/Background: Various tests are used in the screening and diagnosis of Cushing's syndrome (CS). However, none of the existing approaches provide optimal sensitivity and specificity. This study aimed to analyze the accuracy of various tests used for diagnosing CS, including free cortisol measurements in urine, dexamethasone suppression tests (DSTs; 1, 2, and 8 mg), and nocturnal salivary cortisol measurements. The study aimed to identify the most accurate tests using advanced machine learning (ML) techniques. **Methods:** We performed binary classifications using data from 278 patients with CS and 220 controls without CS. We developed five ML classification models, namely, a generalized linear model, decision tree, random forest, naive Bayes, and support vector machine, and compared them using standard performance metrics. We employed Boruta feature selection to identify the most informative clinical and biochemical predictors. **Results:** The naive Bayes model achieved the highest predictive performance, with 100% classification accuracy and a kappa value of 1.0000, followed by the random forest and decision tree models, each with 99.01% accuracy. Boruta feature selection ranked as the most informative predictor post 2-mg DST cortisol, followed by midnight cortisol, post 8-mg DST cortisol, urinary free cortisol, and post 1-mg DST cortisol. The decision tree model identified a cohort-specific post 2-mg DST cortisol threshold of approximately 1.78 µg/dL for distinguishing CS-positive from CS-negative cases. **Conclusions:** ML models, particularly naive Bayes, exhibited strong internal performance in distinguishing patients with CS from non-CS controls using routinely assessed biochemical parameters. The prominence of the post 2-mg DST cortisol predictor in the Boruta ranking and decision tree analysis indicates that this variable contributed substantially to the model-based classification in this cohort. However, the identified 1.78 µg/dL threshold should be interpreted as a cohort-specific, data-driven finding rather than as a clinical cutoff recommendation, and external validation in larger multicenter cohorts is required before clinical implementation.

Keywords: cushing's syndrome; machine learning; classification algorithms; performance comparison; medical diagnosis

1. Introduction

Cushing's syndrome (CS) is a clinical condition characterized by various signs and symptoms resulting from chronic exposure to high cortisol levels, whether induced exogenously or endogenously [1]. The annual CS incidence ranges from 1.8 to 3.2 per million people [2]. Mortality rates in patients with CS are higher than in the general population, and even after successful treatment, survival rates may remain low [3]. CS is a severe and chronic disease caused by prolonged tissue exposure to elevated corticosteroid levels. The most common cause of CS is the use of exogenous corticosteroids. In contrast, endogenous CS is a rare condition characterized by cortisol hypersecretion, which can result from various causes [4]. Chronic hypercortisolism is associated with several morbidities, and if left untreated, mortality increases significantly [5].

CS can present with highly nonspecific symptoms such as obesity, depression, hirsutism, and menstrual irregularities. The symptoms also include proximal muscle weakness, plethoric appearance, increased abdominal and facial fat, thinning of the extremities, wide purple striae, easy bruising without significant trauma, and supraclavicular fullness. Despite the wide range of potential symptoms, no symptom is pathognomonic for CS. Therefore, when clinical suspicion arises, biochemical tests are essential to confirm the diagnosis. The primary diagnostic tests for investigating hypercortisolism include the 1-mg dexamethasone suppression test (DST), 24-h urine free cortisol (ufc), midnight salivary cortisol (saliva), and midnight serum cortisol (mc) measurements. After a diagnosis is established, further testing is necessary to determine the underlying cause [6].

Considering the diverse symptoms, findings, and comorbidities as well as the higher mortality rates associated with CS, this complex disease must be considered carefully in the differential diagnosis [7]. Despite the availability of several tests in screening and diagnosis, no test has shown optimal results because of issues such as variable sensitivity and specificity for each patient; high costs; the need for medication, additional consultations, or even hospitalization. These challenges are particularly evident in primary and secondary care centers [8,9]. None of the first-line tests have shown ideal sensitivity or specificity, necessitating the search for a simple, measurable, and optimal diagnostic test [10].

The use of machine learning (ML) frameworks has addressed various diagnostic and prognostic challenges in the clinical context of CS [11]. In this domain, automated systems have been implemented for analyzing urinary steroid profiles, which assist in distinguishing healthy metabolic states and abnormal patterns associated with CS [12]. Additionally, ML-based classification approaches are effective when applied to molecular datasets, specifically the gene expression profiles of tumor tissues [13]. In addition to initial detection, these computational tools have been crucial in identifying key variables that predict both immediate and long-term clinical outcomes following surgical interventions for CS [14]. Furthermore, novel ML applications have been developed to detect the facial dysmorphisms, which are characteristic of endocrine disorders, thereby enhancing the efficiency of diagnosis and patient monitoring [15]. Our current approach builds upon these foundations by focusing on high-precision screening through stable biochemical feature selection [16].

The primary goal of the present study was to develop a comprehensive diagnostic decision-support system rather than a primary first-line screening tool. By integrating both traditional screening markers (1-mg DST, ufc) and etiological indicators such as adrenocorticotropic hormone (ACTH, 8-mg DST), the model is designed to facilitate the rapid and objective validation of CS diagnosis in clinical settings. Because the model uses parameters inherent to the clinical inclusion criteria, its high performance reflects its ability to optimize mathematically and standardize existing diagnostic protocols rather than identifying independent predictive variables.

Building on our previous investigation using ML in endocrine diagnostics [16], the present study aimed to evaluate the potential of ML-based classification models to support the diagnosis of CS using routinely assessed clinical and biochemical parameters. Specifically, we sought to compare the diagnostic performance of various classification algorithms and to provide a standardized, data-driven framework that may assist clinicians in the objective assessment of CS.

2. Methods

2.1 Study Design and Population

The clinical cohort utilized in this study has been described previously in our preliminary research [16]. While the patient population remains the same, the current study expands significantly upon the initial findings by implementing advanced ML architectures-including the naive Bayes, random forest, and support vector machine (SVM) models-and by utilizing robust feature selection techniques such as Boruta to achieve higher diagnostic precision and interpretability.

In this study, we analyzed the retrospective medical records of 498 unique patients (319 women and 179 men, with a mean age of 52.02 ± 13.33 years). These records were for patients who presented to the Akdeniz University Faculty of Medicine Endocrinology outpatient clinic between 2014 and 2023 due to clinical suspicion of CS or incidental adrenal adenomas. The inclusion and exclusion criteria were as follows:

● CS+ group: Patients with biochemically confirmed endogenous hypercortisolism, defined by at least two positive screening tests (e.g., 1-mg DST >1.8 µg/dL, elevated 24-h ufc, or abnormal midnight salivary cortisol) and final clinical diagnosis.

● CS− (Control) group: Included patients with nonfunctional adrenal adenomas who demonstrated normal biochemical screening results and no clinical signs of CS. This control group included a total of 220 Cushing-negative patients with nonfunctioning adrenal adenomas.

● Exclusion criteria: Patients with exogenous glucocorticoid use, incomplete core diagnostic data, or those under 18 years of age were excluded. For patients with multiple visits, only the data from the initial diagnostic workup were included to ensure the independence of the observations.

2.2 Dataset Profile

The CS dataset comprised 498 samples and 10 candidate predictor variables including biochemical and demographic parameters. The clinical and biochemical characteristics of the cohort are summarized in Table 1. We included the (adrenal and pituitary) imaging reported in Table 1 to characterize the cohort descriptively; however, they were not included as candidate predictors in the ML models, as our primary objective was to evaluate the biochemical screening markers. The target variable represents four subtype classes: adrenal Cushing (AC), pituitary Cushing (PC), subclinical Cushing's syndrome (SC), and ectopic Cushing (EC) for patients with CS. It also represents nonfunctional adrenal adenoma in patients without CS (CS-). Although the dataset includes five distinct labels (CS−, AC, PC, EC, and SC), they were aggregated into a binary format (CS+ vs. CS-) in this study. An important methodological consideration is that we have based the definition of the CS+ outcome, in part, on biochemical screening indicators such

Table 1. Characteristics of patients at the time of diagnosis.

Features	Total (n = 498)	CS+ (n = 278)	CS− (n = 220)
Age (years)	52.02 ± 13.33	50.45 ± 14.20	54.01 ± 11.95
Sex (female/male)	319/179	185/93	134/86
Biochemical Tests			
Basal cortisol (µg/dL)	16.85 ± 8.42	21.30 ± 9.12	11.22 ± 4.35
Basal ACTH (pg/mL)	31.20 (14.5–45.8)	38.40 (18.2–55.4)	24.10 (12.4–32.6)
1-mg DST (µg/dL)	8.45 ± 9.15	14.25 ± 10.30	1.10 ± 0.45
2-mg DST (µg/dL)	7.15 ± 8.20	12.35 ± 9.45	0.65 ± 0.25
8-mg DST (µg/dL)	4.25 ± 5.10	6.15 ± 6.45	0.85 ± 0.35
Midnight cortisol (µg/dL)	12.45 ± 10.12	19.35 ± 11.45	3.75 ± 2.10
Night salivary cortisol (µg/dL)	0.42 (0.15–0.95)	0.78 (0.45–1.42)	0.12 (0.08–0.18)
24 h urine cortisol (µg/24 h)	185.4 (42.6–415.2)	354.2 (112.4–658.2)	48.6 (24.2–68.4)
Imaging performed (n, %)			
CT adrenal imaging (n, %)	184, 36.9%	132, 47.5%	52, 23.6%
CT/MRI pituitary imaging (n, %)	115, 23.1%	98, 35.3%	17, 7.7%

ACTH, adrenocorticotropic hormone; CT, computed tomography; DST, dexamethasone suppression test; MRI, magnetic resonance imaging; CS+, Positive for Cushing's syndrome; CS-, Negative for Cushing's syndrome.

as the 1-mg DST, ufc, and mc, which were also included as candidate input features in the development of ML models. Therefore, the present modeling framework should be interpreted as a data-driven decision-support approach for standardizing and optimizing existing diagnostic information rather than a fully independent diagnostic discovery model. This overlap between outcome-defining clinical criteria and model predictors may have contributed to the very high classification performance observed in this study and represents a potential source of incorporation bias. Accordingly, the reported performance metrics should be interpreted within this clinical and methodological context. We made this decision to focus the analysis on the primary diagnostic distinction between CS+ and CS- cases and to evaluate whether routinely assessed clinical and biochemical parameters could support this clinically relevant binary classification task. The descriptive statistics of the clinical cohort, including means, standard deviations, and proportional distributions, are presented in Table 1. Analysis of the dataset revealed a nonuniform distribution of missing values, primarily because diagnostic evaluations were tailored to individual patient presentations and clinician discretion. These missing value rates (MVR%) reflect those in real-world clinical practice, in which certain tests are prioritized based on emergent symptoms. Furthermore, several feature distributions exhibited significant skewness, necessitating robust analytical approaches. The dataset is characterized by a moderate class distribution difference, with the CS+ group constituting a larger proportion (n = 278) of the sample than the CS- (control group) (n = 220).

2.3 Preprocessing

Before developing the ML models, preprocessing focused on three main issues: missing data, outlying biochemical values, and skewed feature distributions. We handled missing values using the imputation strategy described below, while we reduced the skewness of the biochemical variables and minimized the influence of extreme values by applying a log10 transformation [16].

From the total dataset of 498 samples, was randomly allocated for training approximately 80% (n = 397) and cross-validation, while a separate, independent, hold-out test set consisting of 101 instances (approximately 20%) was reserved for final performance evaluation.

In creating the classification models, we applied 10-fold cross-validation to the training dataset. We employed stratified sampling during the 10-fold cross-validation procedure to preserve the proportional distribution of CS+ and CS− cases within each fold. Given the moderate class imbalance in the dataset (278 CS+ vs. 220 CS-), this approach ensured stable training dynamics and reduced the risk of biased performance estimation. We used cross-validation to assess model consistency across different subsets of the training data, improve its generalizability, and minimize the risk of overfitting [17]. For each algorithm, we recorded the mean 10-fold cross-validation accuracy obtained from the training set as an internal estimate of model stability. We then evaluated the final performance of the model using the independent hold-out test set, which we consider to be the primary performance estimates.

2.4 Missing Value Imputation

In clinical practice, diagnostic protocols for CS are often individualized, resulting in a nonuniform distribution of biochemical tests. Physicians may defer specific evaluations if prior results yield conclusive diagnostic insights. For example, 8-mg DST is typically reserved for etiologic differentiation following initial screening using 1-mg DST.

In cases involving suspected nonfunctioning adrenal adenomas, 8-mg DST is frequently bypassed. In contrast, based on the severity of clinical manifestations, clinicians may prioritize high-dose suppression tests over standard low-dose protocols. To reflect these real-world clinical dynamics and prevent the loss of valuable diagnostic information, we retained observations with incomplete data rather than performing listwise deletion. Considering the pronounced right-skewed distributions and the presence of extreme biochemical values, we selected median imputation as a robust and distribution-resistant approach. Unlike mean imputation, the median is less sensitive to outliers, which is particularly important in endocrine datasets with high biological variability. More complex techniques such as K-nearest neighbor imputation may introduce additional variance and parameter sensitivity, whereas multiple imputation, although theoretically robust, requires additional distributional assumptions and may complicate interalgorithm comparisons in this moderate-sized cohort. Therefore, a conservative and stable median-based strategy was deemed most appropriate for this moderate-sized clinical cohort. However, we acknowledge that single-value imputation methods, including median imputation, may reduce variance and underestimate uncertainty compared with multiple imputation approaches. Therefore, we used median imputation for the primary model comparison to ensure a consistent preprocessing pipeline across all algorithms, while we applied multivariate imputation by chained equations (MICE) as a sensitivity analysis to evaluate the robustness of feature importance rankings.

Addressing missingness in clinical datasets remains a complex statistical endeavor, as no universal imputation strategy is optimally suited for every analytical scenario [18]. Although removing cases with missing values is a common approach, we deemed it to be inappropriate for this study due to the high missing value rate (MVR%) (ranging from 0% [age, sex, and saliva] to 81.93% [8-mg DST]) in the original cohort. Such an exclusion would have resulted in a significant reduction in sample size and compromised the statistical power of the models. To maintain strict methodological rigor and prevent data leakage, we managed missing values for all the biochemical and clinical features using a median imputation strategy. Specifically, we calculated the median values for each feature based solely on the training dataset. We then applied these derived constants consistently to fill in missing entries in both the validation and the independent hold-out test sets. This protocol ensured that both of these datasets remained completely unseen during model development, thus minimizing information leakage and enabling an unbiased assessment of model performance [19].

In addition to the primary median imputation strategy, we performed a sensitivity analysis using MICE to evaluate the robustness of the feature importance rankings against alternative imputation frameworks. We applied the MICE procedure exclusively to assess comparative stability and not for the final model training or performance reporting. This additional step ensured that the key predictors we identified were not artifacts of a specific imputation approach.

2.5 Asymmetric Distribution of Clinical Features

Among the classifiers used in this study, NB assumes that continuous features follow a Gaussian (normal) distribution, while the SVM is sensitive to feature scaling and distributional properties. When these assumptions are violated, predictive performance may be adversely affected [20]. To address this issue, we applied a log10 transformation to normalize the highly skewed biochemical features (Table 2), thereby enhancing model stability and ensuring more reliable classification performance across all algorithms. In contrast, we expected classifiers that do not rely on specific distributional assumptions to maintain robust performance across varied data structures, provided that class separations remain distinct [21]. We subsequently evaluated and reported the extent of remaining skewness using the Fisher–Pearson skewness coefficient [22].

2.6 Feature Selection Method

The feature selection method plays a crucial role in predicting the impact of features in datasets with large numbers of variables on the performance of an ML model [23]. Removing irrelevant features both simplifies the dataset and enhances the interpretability of the model [24]. Among various possible feature selection methods, we selected the Boruta algorithm for this study due to its robust all-relevant selection approach [25]. Unlike traditional heuristic methods that seek a minimal optimal subset of features, Boruta identifies all the statistically significant variables that are associated with the target variable. It achieves this by creating shadow features, which are randomly permuted versions of the original attributes, to establish a rigorous statistical baseline for assessing feature importance, thereby ensuring that only features with higher importance than random noise are retained for model development. To evaluate the stability of feature selection, we performed a bootstrap resampling procedure with 50 iterations. For each bootstrap sample, we repeated the feature selection process and computed a stability index to measure the frequency and consistency with which the key predictors were retained across resampled datasets.

2.7 ML Classification and Evaluation Methods

The literature contains a wide range of ML algorithms, which are generally categorized into classification, regression, and clustering algorithms. While some algorithms are specific to one category, others can be applied across multiple categories.

A generalized linear model (GLM) is an ML algorithm that attempts to determine the distribution of the dependent variable using a linear function of the independent variables

Table 2. Skewness coefficients of clinical and biochemical features before and after log_{10} transformation.

Feature	Skewness (before log)	Skewness (after log)
Night salivary cortisol (saliva)	15.96	0.80
Basal ACTH (bacth)	9.07	−0.04
Urinary free cortisol (ufc)	7.28	0.60
8-mg DST (mg8)	5.09	1.35
2-mg DST (mg2)	3.88	0.61
1-mg DST (mg1)	3.71	0.45
Midnight cortisol (mc)	3.16	0.44
Basal cortisol (bc)	2.56	0.67

ACTH, adrenocorticotropic hormone; DST, dexamethasone suppression test.

[26]. Because it is easy to interpret the parameters produced by this model, we chose it to solve the CS-modeling problem. This algorithm has been successfully applied in healthcare settings [27]. Although GLM was previously used to develop a mobile diagnostic application with high accuracy [16], it is used here as a statistical baseline to measure the improvement provided by more complex algorithms.

The decision tree (DT) algorithm is a versatile, flexible, and interpretable ML algorithm that simplifies complexity by dividing the data into meaningful segments. Its ability to produce a simplified tree structure and provide an easily interpretable visual output makes it a popular choice in medical applications [28]. We also adopted an SVM because it is effective in handling complex decision boundaries [21], while we selected random forest (RF) for its robustness against overfitting [29]. Finally, we included the NB algorithm because its high computational efficiency, and demonstrated effectiveness even with relatively small clinical datasets [30].

2.8 Evaluation Metrics

We evaluated the performance of each model using the data reserved for testing purposes. In the field of classification, the metrics used most commonly for evaluating model performance include accuracy, sensitivity, specificity, expected accuracy (EA), and kappa [31]. These metrics were evaluated following the methods of Aydemir et al. [16].

To express these evaluation metrics mathematically, we used the confusion matrix (CM), which indicates how well a given model differentiates between actual and predicted classes (Table 3). The key terms in this table are true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

We assessed the performance and reliability of the models developed using robust evaluation metrics, including accuracy and kappa. Accuracy is defined as the proportion of correctly classified instances among all evaluated cases; it is calculated as Accuracy = (TP + TN)/(TP + TN + FP + FN). Cohen's kappa coefficient (κ) quantifies the degree of agreement between predicted and actual classifications beyond chance expectation. It is calculated as $\kappa = (Po - Pe)/(1 - Pe)$, where Po denotes the observed agreement (Accuracy) and Pe denotes the agreement expected by chance. Kappa values range from 0 to 1, with higher values indicating stronger agreement beyond chance.

Table 3. Confusion matrix.

Classification		Actual	
		CS+	CS−
Prediction	CS+	TP	FP
	CS−	FN	TN

CS+, Positive for Cushing's syndrome; CS-, Negative for Cushing's syndrome; TP, true positive; FP, false positive; FN, false negative; TN, true negative.

In addition to accuracy and kappa, we used several other metrics to provide a more robust evaluation, particularly given the moderate class imbalance in the dataset (278 CS+ vs. 220 CS-). These included:

- Balanced accuracy: The average of sensitivity and specificity, this metric is given by Balanced Accuracy = [TP/(TP+FN) + TN/(TN+FP)] / 2. It is less sensitive to class imbalance than standard accuracy.
- Precision: The proportion of predicted positive cases that are truly positive, it is given by Precision = TP/(TP + FP).
- Recall (Sensitivity): The proportion of actual positive cases that are identified correctly, it is calculated as Recall = TP/(TP + FN).
- F1-score: This metric is the harmonic mean of Precision and Recall, and it is given by F1 = 2 × (Precision × Recall)/(Precision + Recall).
- ROC–AUC (Receiver Operating Characteristic–Area Under the Curve): This is a threshold-independent metric that quantifies the ability of a model to distinguish between the classes CS+ and CS-. An AUC of 1.0 indicates perfect separation, while AUC = 0.5 indicates random guessing [31].

These metrics were calculated for all models along with the previously reported accuracy measures from Table 4.

Table 4. Performance evaluation metrics of ML models calculated for the independent hold-out test set (n = 101).

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score	Expected accuracy	Kappa
NB	100.00%	100.00%	100.00%	1.0000	0.5004	1.0000
RF	99.01%	100.00%	97.96%	0.9905	0.5007	0.9802
DT	99.01%	100.00%	97.96%	0.9905	0.5007	0.9802
SVM	98.02%	98.08%	97.96%	0.9808	0.5004	0.9604
GLM	97.03%	98.08%	95.92%	0.9714	0.5007	0.9405

ML, machine learning; NB, naive Bayes; RF, random forest; DT, decision tree; SVM, support vector machine; GLM, generalized linear model.

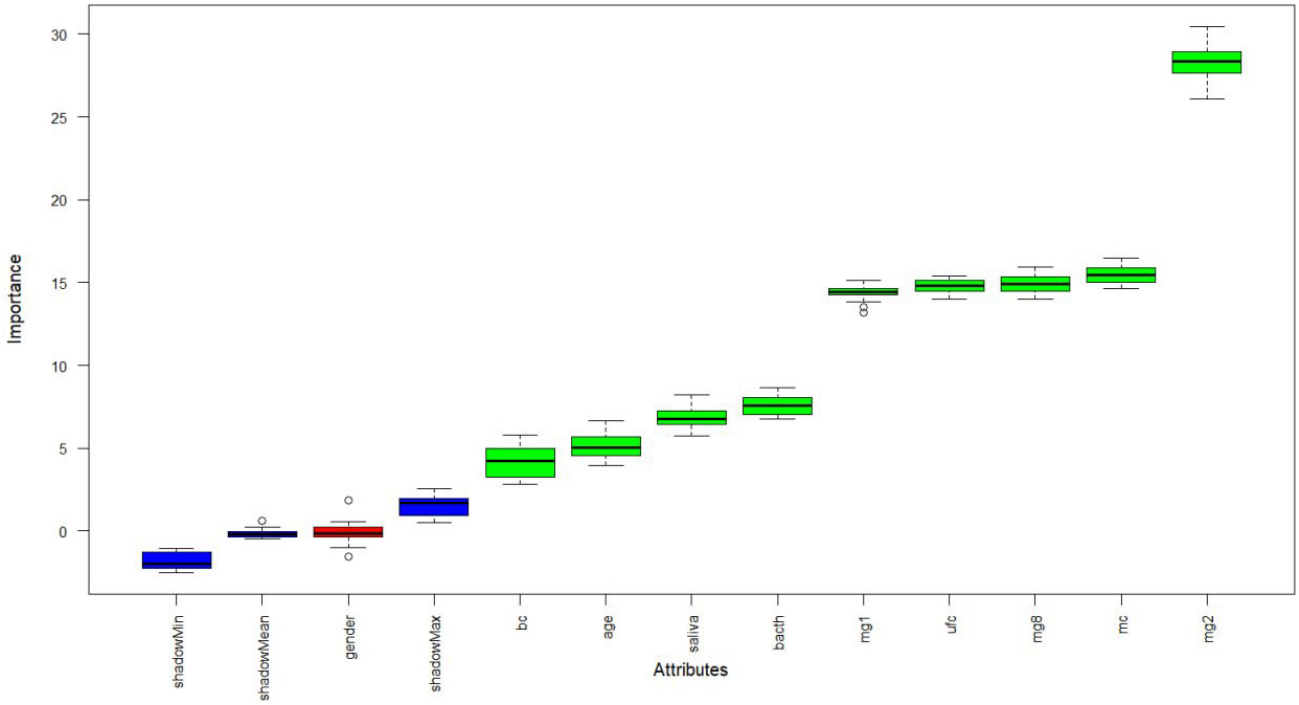

Fig. 1. Feature importance assessment of biochemical and clinical variables for CS detection using the Boruta algorithm. Abbreviations: bc, basal cortisol; CS, Cushing's syndrome; saliva, night salivary cortisol; bacth, basal adrenocorticotropic hormone; mg1, cortisol after 1-mg dexamethasone suppression test (DST); mg2, cortisol after 2-mg DST; mg8, cortisol after 8-mg DST; ufc, urinary free cortisol; mc, midnight cortisol.

2.9 Subgroup Analysis

To determine whether performance of a given model remained consistent across demographic strata, we performed subgroup analyses with respect to gender and age. We stratified the patients by gender and by age, using the mean age of the cohort as the cutoff value between “young” and “old”.

3. Results

In this study, the Boruta algorithm was used to determine the importance ranking of the features in the dataset, and the results are presented in Fig. 1.

To ensure the reliability of the feature selection process, the stability of the Boruta algorithm was evaluated using 50 bootstrap iterations. For each bootstrap sample, the Boruta procedure was repeated, and a stability index was calculated as the proportion of iterations in which each predictor was retained as important. The ranking of the leading predictors remained unchanged when median imputation was compared with MICE. Post 2-mg DST cortisol, post 1-mg DST cortisol, and midnight cortisol each showed a stability index of >0.95. The remaining retained biochemical predictors also demonstrated stable selection frequencies, including post 8-mg DST cortisol and ufc. These findings indicate that the Boruta feature ranking was not driven by a single data partition or by a specific imputation strategy.

The results obtained from the Boruta algorithm are presented in Fig. 1. In this figure, the horizontal axis shows the clinical features along with three synthetic benchmarks generated by the algorithm: shadowMin, shadowMean, and shadowMax. These shadow features serve as a randomized baseline for comparison. Specifically, shadowMax represents the maximum importance achieved by a noninformative shuffled attribute; clinical features that significantly

outperform this shadowMax reference line are confirmed as significant predictors of CS. The vertical axis shows the importance scores assigned to each feature by the algorithm. In this plot, the importance of a feature increases as one moves to the right along the horizontal axis. As shown in the figure, according to the Boruta feature importance ranking, the 2-mg DST (mg2) emerged as the most significant predictor for distinguishing CS cases, with an importance score nearly twice that of the other attributes. The four predictors mc, mg8, ufc, and mg1 were next in the order of importance. Although the 1-mg DST (mg1) remains a cornerstone of international clinical guidelines, its position in our model hierarchy indicates that within this specific cohort, the 2-mg DST predictor showed higher feature importance within this specific dataset and modeling framework, enabling the ML algorithms to achieve high diagnostic separation. These highly ranked biochemical features were subsequently incorporated into the diagnostic models. Following these features in the importance ranking are bacth, saliva, age, and bc. The gender feature, highlighted in red in Fig. 1, shows no significant importance with respect to the dependent variable. Based on these Boruta results, we found gender to be statistically noninformative; therefore, we excluded it from the final modeling phase. Consequently, we constructed the classification models using nine predictive biochemical features.

Using the training data, which constituted 80% of the entire dataset, we developed five different ML classification models. The comparative performance of these five ML models is summarized in Table 4. The NB model achieved the highest performance across all metrics, with 100% sensitivity and specificity, resulting in a perfect kappa coefficient of 1.0. This was closely followed by the RF model, which demonstrated an accuracy of 99.01%. Notably, all models maintained a kappa score of >0.94, indicating "almost perfect" agreement beyond chance. To provide transparency in interpreting the magnitude of agreement beyond chance, we have reported the EA, which represents the agreement expected by chance and constitutes an intrinsic component of Cohen's kappa calculation. The performance metrics presented in Table 4 correspond to the final evaluation results obtained from the independent hold-out test set (n = 101). These values represent single test-set outcomes following model optimization via 10-fold cross-validation on the training data and are not cross-validation averages.

All models maintained a high level of reliability, with the GLM model used as a statistical baseline achieving an accuracy of 97.03% and a kappa value of 0.9405 on the test set.

The CM and statistical results obtained during the testing phase of the models created with the GLM, SVM, DT, RF, and NB algorithms are presented in Table 5, and the diagram generated by the DT model is presented in Fig. 2.

In the binary encoding used to develop the DT model, 0 represents CS- and 1 represents CS+. In Fig. 2, the decimal values within each node are the predicted probabilities for the classes CS- and CS+, respectively. For example, the values 0.99 and 0.01 indicate a high probability of classification as CS-, whereas the values 0.01 and 0.99 indicate a high probability of classification as CS+. The percentage shown below each node represents the proportion of observations assigned to that node.

Table 5. Confusion matrix raw counts for evaluated models (n = 101).

Model	TP	TN	FN	FP
NB	52	49	0	0
RF	52	48	0	1
DT	52	48	0	1
SVM	51	48	1	1
GLM	51	47	1	2

Note: The test set consists of 52 CS+ cases and 49 CS- (NF/Control) cases (total n = 101).

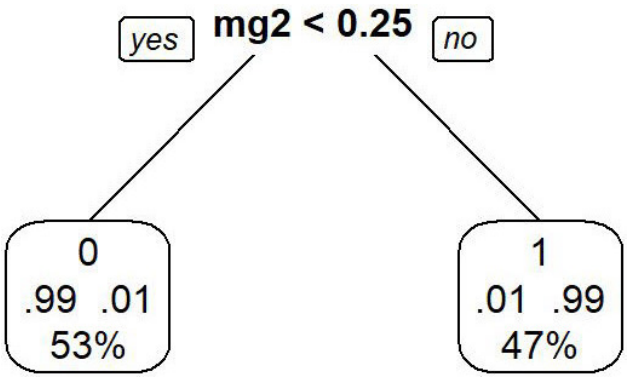

Fig. 2. Decision tree (DT)-derived diagnostic classification model for Cushing's syndrome based on the post 2-mg dexamethasone suppression test cortisol threshold and key biochemical predictors.

Based on the CM results, the GLM model made errors on only 3 of the 101 instances. Accordingly, the accuracy is calculated using the following formula:

$$\text{Accuracy (Acc)} = \frac{TP + TN}{N} = \frac{47 + 51}{47 + 51 + 1 + 2} = 97.03\% \quad (1)$$

where $N = TP + TN + FP + FN$ is the total number of cases.

Similarly, the SVM model made 2 classification errors out of 101 instances. Specifically, one patient from the CS- group was incorrectly predicted to be CS+ (false positive), and one patient from the CS+ group was incorrectly predicted to be CS- (false negative), resulting in an overall classification accuracy of 98.02%. This finding demonstrates that the model correctly classified a high percentage

of instances, showing robust performance in distinguishing the two classes.

For both DT and RF models, only one error was observed in the 101-instance test set. In both models, one patient who was actually CS- was incorrectly classified as CS+ (FP = 1), whereas all CS+ cases were correctly identified (FN = 0), yielding a classification accuracy of 99.01%.

Fig. 2 shows the DT classification using the mg2 feature. To interpret this classification on the original scale, an inverse log10 transformation is required. The split value of 0.25 corresponds to approximately $10^{0.25} \approx 1.78$ µg/dL. Accordingly, patients with a 2 mg DST cortisol value below 1.78 µg/dL are classified as CS- (Control), whereas those with values equal to or above 1.78 µg/dL are classified as CS+. This data driven cutoff should be interpreted as a cohort specific threshold for the 2 mg DST rather than as a direct validation of guideline recommended thresholds established for the 1 mg DST.

The CM also shows that the NB model made no classification errors out of 101 instances. Upon examining CM, all 49 patients who were actually CS- were correctly classified as CS-, and all 52 patients who are actually CS+ were correctly classified as CS+. The overall classification accuracy of this model is 100%, i.e., perfect classification performance.

The high kappa values (0.9405–1.000) listed in Table 4 indicate excellent agreement between predicted and observed classifications.

The alignment between the average cross-validation performance on the training set and the results obtained from the independent test set suggests stable learning behavior and good generalization within this cohort. However, external or temporal validation studies are required to assess potential overfitting and broader clinical applicability more definitively.

To ensure a comprehensive evaluation and account for the class imbalance in the dataset, we calculated additional robust metrics for all models. As summarized in Table 6, all models exhibited high precision and recall, with NB achieving perfect scores across all metrics. The high ROC–AUC values (>0.97 for all models) further highlight the excellent diagnostic separation between CS+ and CS- cases.

Table 6. Consolidated performance metrics of ML models for CS detection.

Model	Balanced Accuracy	Precision	Recall	ROC–AUC
NB	1.000	1.000	1.000	1.000
RF	0.990	0.981	1.000	0.999
DT	0.990	0.981	1.000	0.990
SVM	0.980	0.981	0.981	0.995
GLM	0.970	0.962	0.981	0.993

ROC–AUC, Receiver Operating Characteristic–Area Under the Curve.

The high consistency between the mean 10-fold cross-validation accuracy (99.37%, calculated during model development on the training set) and the test set accuracy range of the evaluated models (97.03%–100%) suggests stable learning behavior across the classification algorithms. However, considering the relatively small size of the independent test set (n = 101), these findings should be interpreted with caution, and external validation is required to assess model generalizability more definitively.

Subgroup Analyses

To evaluate the robustness of the models against demographic variation, we performed subgroup analyses with respect to sex and age. The cohort included 319 female and 179 male participants. For the age-based analysis, patients were stratified into those younger (n = 222) or older (n = 276) than the mean age of the cohort (52 years). Using the independent test set, we further evaluated the NB and RF models with respect to sex. In the test set, 72 patients were classified as female and 29 patients as male, whereas the age-based subgroups included 45 patients younger than 52 years and 56 patients aged ≥52 years. The NB model achieved 100.00% accuracy in all demographic subgroups including female, male, patients younger than 52 years, and patients aged ≥52 years. The RF model also showed stable subgroup performance, with accuracy values of 98.61% for females, 100.00% for males, 97.78% for patients younger than 52 years, and 100.00% for patients aged ≥52 years. These subgroup-specific results indicate that the two best performing models retained high internal classification performance across the examined sex and age strata. However, these findings should be interpreted cautiously because of the limited sample sizes of the subgroups, and external validation using larger and more heterogeneous cohorts is required.

4. Discussion

In this study, based on biochemical test results commonly employed in the diagnosis of CS, we compared the performance of several widely used ML classification algorithms. The high accuracy rates obtained in our analysis indicate that ML models hold substantial potential as clinical decision-support systems in the diagnostic process. However, it is essential to discuss our findings in light of the existing literature while also addressing the methodological challenges and clinical implications of the study.

Recently, the number of studies investigating ML applications in CS diagnostic assessment has been increasing. For instance, Lyu et al. [29] evaluated ML-based approaches in the context of ACTH-dependent hypercortisolism and reported a ROC-AUC value of 0.976 ± 0.03. Similarly, Golounina et al. [32] conducted a multialgorithmic comparison and reported a ROC-AUC of 0.920, with a sensitivity of 77.8% and specificity of 97.1%. Collectively, these studies demonstrate that ML techniques can support diagnostic decision-making in CS-related clinical contexts.

In our study, the NB algorithm achieved an accuracy of 100% with a kappa value of 1.0000, whereas RF and DT achieved 99.01% accuracy. These values are higher than those reported in previous studies on ML-based CS. This difference may be attributed to several methodological and clinical factors. First, we designed the present study as a binary classification task (CS+ vs. CS-), whereas some previous studies have addressed more complex diagnostic scenarios [32,33]. Second, the biochemical variables used as predictors, such as DST results, ufc, and cortisol measurements, are strongly related to the clinical criteria used for defining CS status, which may have increased class separability. Third, the retrospective single-center cohort and the relatively limited size of the independent test set may have contributed to the optimistic performance estimates. Therefore, although the observed performance indicates strong internal classification ability within this cohort, the superiority of our results over those of previous studies should be interpreted with caution until external validation is performed.

Because the present analysis was limited to the binary classification into CS+ or CS, the models were not designed or trained to distinguish various CS etiologies. Consequently, etiology-level diagnostic interpretation was beyond the scope of the current analysis. This limitation should be addressed in future studies using appropriately designed multiclass classification models with sufficient sample sizes for each diagnostic category.

The DT model in our study identified an approximate threshold of 1.78 µg/dL for post 2-mg DST cortisol levels to distinguish between CS+ and CS- cases. However, this threshold should not be equated with the guideline-recommended 1.8 µg/dL cutoff, which applies specifically to the 1-mg DST test. Although the numerical similarity is noteworthy, the two tests involve different doses of dexamethasone and therefore cannot be interpreted as being directly interchangeable. Accordingly, the 2-mg DST threshold identified in this study should be regarded as a data-driven, cohort-specific finding that requires further validation before any clinical threshold recommendation can be made.

The feature importance analysis performed using the Boruta algorithm identified post 2-mg DST cortisol levels as the most influential variable, followed by midnight cortisol, post 8-mg DST cortisol, and 24-h ufc levels. This finding supports the notion that classical biochemical tests remain strong predictors from an ML perspective. The ranked feature importance hierarchy we have presented not only enhances model interpretability but also provides clinically relevant insight as to which biochemical markers contribute most strongly to the binary distinction between CS+ and CS- cases. From a practical standpoint, this underscores the importance of test selection, timing, and quality control in building robust ML-based diagnostic decision-support frameworks. We also acknowledge that missing data in clinical registries such as a clinician's decision to bypass specific confirmatory tests based on initial results may carry inherent clinical information. Although we used median imputation to preserve the sample size for model comparisons, this approach may mask the clinical reasoning associated with the missing data. This is a significant limitation of the present work, and we therefore suggest that future frameworks should explore missing-indicator methods to capture these subtle clinical cues.

The high accuracy achieved across all the models indicates that ML algorithms may serve as valuable adjuncts to traditional diagnostic methods such as the 1-mg DST, ufc, and late-night salivary cortisol tests. The ML-based systems offer the advantage of rapid interpretation of the data, facilitating the integration of predictive outputs into clinical workflows. Consistent with this, previous studies have emphasized that ML models may serve as noninvasive and cost-effective alternatives to traditional invasive diagnostic procedures [16,34].

Our finding that the sex variable did not contribute significantly to model performance further indicates that biochemical parameters can predict CS reliably, irrespective of the gender of the patient. Thus, a standardized biochemical protocol may be applicable across all patients; however, this observation also requires further validation in larger and more heterogeneous populations.

Despite the high accuracy levels achieved, we acknowledge several limitations of this study. First, measurement conditions, test timing, laboratory variability, and missing data may have influenced model performance. Second, although the classification algorithms demonstrated excellent accuracy, this study did not address causal inference or temporal progression of the disease; therefore, the predictive scope of the models remains limited to the existing test results rather than constituting an evaluation of longitudinal disease dynamics. Third, there was a partial overlap between the biochemical criteria used to define CS+ status and the predictor variables used in model development. This overlap may have inflated the apparent diagnostic performance of the models. Our results should therefore be interpreted as reflecting the ability of ML algorithms to standardize established diagnostic patterns rather than to identify CS using entirely independent predictors.

Because the dataset we used was derived from a single-center cohort, external validity and generalizability of our results may be limited. Similar concerns have been raised about previous studies. For example, Zoli et al. [33] reported postoperative-outcome prediction accuracies ranging from 81% to 100% in patients with CD using ML, and they highlighted the need for larger and externally validated datasets. Further, because measurement conditions, test timing, laboratory variability, and missing data may have influenced model performance, it is essential to ensure consistent data quality for reliable model deployment in real-world clinical environments. In addition, although

the classification algorithms we used demonstrated excellent accuracy, we did not address causal inference or temporal progression of the disease in this study; the predictive scope of the models therefore remain limited to the existing test results.

Based on these considerations, we recommend that future research should pursue the following directions:

• Conduct multicenter studies that include diverse geographic and demographic populations to enable robust external validation.

• Investigate clinical workflow integration such as compatibility with laboratory information systems and electronic health records to assess real-world feasibility.

• Incorporate multimodal data including imaging findings, genetic/omics profiles, and additional clinical variables to further enhance model performance and generalizability.

• Develop and evaluate explainable artificial intelligence frameworks to address the black-box problem and improve clinician acceptance.

• Explore time series and longitudinal data to model disease progression and predict subtype evolution over time.

By moving beyond preliminary assessments and implementing a more sophisticated computational architecture for CS detection, this study has advanced our previous research framework [16] significantly. Although our earlier work established the feasibility of using ML in a clinical context, the current study introduces a broader suite of high-performance algorithms including NB, RF, and SVM to ensure a more comprehensive performance comparison. Furthermore, by integrating the Boruta feature selection protocol, we achieved greater stability in identifying the most discriminative biochemical markers such as the 1-mg and 2-mg DST tests. By addressing the limitations observed in initial modeling attempts, this methodological evolution improves diagnostic precision and provides a more transparent and interpretable model for clinical decision-making. Using the same clinical cohort with the aforementioned advanced techniques, we demonstrated that refined algorithmic approaches can extract deeper diagnostic insights and offer a more reliable screening tool for endocrine disorders.

5. Conclusions

In this study, we demonstrated that ML algorithms, particularly NB, RF, and DT, can achieve high diagnostic accuracy in detecting CS using standard biochemical tests. The identified 2-mg DST threshold of 1.78 µg/dL represents a cohort-specific, data-driven finding; however, it should not be interpreted as direct validation of guideline thresholds established for the 1-mg DST. Although these models exhibit strong internal validity, they should be positioned as auxiliary decision-support tools to assist clinicians in the objective validation of hypercortisolism. External validation in multicenter, prospective cohorts is essential before broader clinical implementation.

Key Points

• This study developed a comprehensive ML-based decision-support system to improve the diagnostic accuracy of CS.

• The Boruta algorithm identified 2-mg DST and midnight cortisol as the most significant biochemical predictors.

• The naive Bayes model achieved 100% accuracy on an independent test set.

• A cohort-specific, data-driven threshold of 1.78 µg/dL was identified for 2-mg DST.

• This framework can serve as a standardized decision-support tool for the objective diagnosis of hypercortisolism; however, external validation is still required.

• External validation in multicenter prospective studies must ensure broader clinical generalizability.

Availability of Data and Materials

The datasets used or analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions

MA contributed to the clinical conception and design of the study, identification of the study population, endocrine evaluation, diagnostic classification of Cushing's syndrome status, interpretation of biochemical tests, and clinical interpretation of the findings. MC and OO contributed to data preprocessing, missing data handling, machine learning model development, implementation of classification algorithms, performance evaluation, subgroup analysis, and the interpretation of computational results in the context of medical diagnostics. MC, OO, MY, and MA contributed to the study design, methodological planning, feature selection strategy, interpretation of machine learning outputs, academic writing, critical revision of the manuscript, and integration of clinical and computational findings. MA and MY prepared the initial draft of the manuscript. MA revised the manuscript critically to ensure clinical accuracy, endocrine relevance, and diagnostic interpretation. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work, including the accuracy and integrity of the clinical data, analytical procedures, interpretation of results, and final conclusions.

Ethics Approval and Consent to Participate

This study was conducted following the ethical principles outlined in the Declaration of Helsinki and was approved by the Ethics Committee of the Akdeniz University Faculty of Medicine (Approval No.: TBAEK-250, date:

April 25, 2024). Written consent was obtained directly from the participants.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used AI-assisted language tools (ChatGPT and Gemini) only to improve grammar, sentence clarity, readability, and overall textual flow. These tools were not used to generate scientific hypotheses, design the study, collect or manipulate data, perform statistical or machine learning analyses, select references, interpret clinical findings, or draw scientific conclusions. All AI-assisted revisions were critically reviewed, edited, and approved by the authors.

References

[1] Wagner-Bartak NA, Baiomy A, Habra MA, Mukhi SV, Morani AC, Korivi BR, et al. Cushing Syndrome: Diagnostic Workup and Imaging Features, With Clinical and Pathologic Correlation. AJR. American Journal of Roentgenology. 2017; 209: 19–32. https://doi.org/10.2214/AJR.16.17290

[2] Hakami OA, Ahmed S, Karavitaki N. Epidemiology and mortality of Cushing's syndrome. Best Practice & Research. Clinical Endocrinology & Metabolism. 2021; 35: 101521. https://doi.org/10.1016/j.beem.2021.101521

[3] Wengander S, Trimpou P, Papakokkinou E, Ragnarsson O. The incidence of endogenous Cushing's syndrome in the modern era. Clinical Endocrinology. 2019; 91: 263–270. https://doi.org/10.1111/cen.14014

[4] Newell-Price J, Bertagna X, Grossman AB, Nieman LK. Cushing's syndrome. Lancet (London, England). 2006; 367: 1605–1617. https://doi.org/10.1016/S0140-6736(06)68699-6

[5] Lacroix A, Feelders RA, Stratakis CA, Nieman LK. Cushing's syndrome. Lancet (London, England). 2015; 386: 913–927. https://doi.org/10.1016/S0140-6736(14)61375-1

[6] Nieman LK, Biller BMK, Findling JW, Newell-Price J, Savage MO, Stewart PM, et al. The diagnosis of Cushing's syndrome: an Endocrine Society Clinical Practice Guideline. The Journal of Clinical Endocrinology and Metabolism. 2008; 93: 1526–1540. https://doi.org/10.1210/jc.2008-0125

[7] Nieman LK. Cushing's syndrome: update on signs, symptoms and biochemical screening. European Journal of Endocrinology. 2015; 173: M33–8. https://doi.org/10.1530/EJE-15-0464

[8] Baid SK, Rubino D, Sinaii N, Ramsey S, Frank A, Nieman LK. Specificity of screening tests for Cushing's syndrome in an overweight and obese population. The Journal of Clinical Endocrinology and Metabolism. 2009; 94: 3857–3864. https://doi.org/10.1210/jc.2008-2766

[9] Yorke E, Atiase Y, Akpalu J, Sarfo-Kantanka O. Screening for Cushing Syndrome at the Primary Care Level: What Every General Practitioner Must Know. International Journal of Endocrinology. 2017; 2017: 1547358. https://doi.org/10.1155/2017/1547358

[10] Kapoor N, Job V, Jayaseelan L, Rajaratnam S. Spot urine cortisol-creatinine ratio - A useful screening test in the diagnosis of Cushing's syndrome. Indian Journal of Endocrinology and Metabolism. 2012; 16: S376–S377. https://doi.org/10.4103/2230-8210.104099

[11] Laws ER, Catalino MP. Editorial. Machine learning and artificial intelligence applied to the diagnosis and management of Cushing disease. Neurosurgical Focus. 2020; 48: E6. https://doi.org/10.3171/2020.3.FOCUS20213

[12] Wilkes EH, Rumsby G, Woodward GM. Using Machine Learning to Aid the Interpretation of Urine Steroid Profiles. Clinical Chemistry. 2018; 64: 1586–1595. https://doi.org/10.1373/clinchem.2018.292201

[13] Yang JY, Yang MQ, Luo Z, Ma Y, Li J, Deng Y, et al. A hybrid machine learning-based method for classifying the Cushing's Syndrome with comorbid adrenocortical lesions. BMC Genomics. 2008; 9: S23. https://doi.org/10.1186/1471-2164-9-S1-S23

[14] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. Journal of Medical Internet Research. 2016; 18: e323. https://doi.org/10.2196/jmir.5870

[15] Wei R, Jiang C, Gao J, Xu P, Zhang D, Sun Z, et al. Deep-Learning Approach to Automatic Identification of Facial Anomalies in Endocrine Disorders. Neuroendocrinology. 2020; 110: 328–337. https://doi.org/10.1159/000502211

[16] Aydemir M, Çakir M, Oral O, Yilmaz M. Diagnosis of Cushing's syndrome with generalized linear model and development of mobile application. Medicine. 2025; 104: e42910. https://doi.org/10.1097/MD.0000000000042910

[17] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019; 14: e0224365. https://doi.org/10.1371/journal.pone.0224365

[18] Stavseth MR, Clausen T, Røislien J. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Medicine. 2019; 7: 2050312118822912. https://doi.org/10.1177/2050312118822912

[19] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019; 25: 44–56. https://doi.org/10.1038/s41591-018-0300-7

[20] Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of Biomedical Informatics. 2019; 90: 103089. https://doi.org/10.1016/j.jbi.2018.12.003

[21] Isci S, Kalender DSY, Bayraktar F, Yaman A. Machine Learning Models for Classification of Cushing's Syndrome Using Retrospective Data. IEEE Journal of Biomedical and Health Informatics. 2021; 25: 3153–3162. https://doi.org/10.1109/JBHI.2021.3054592

[22] Shi R, McLarty JW. Descriptive statistics. Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology. 2009; 103: S9–S14. https://doi.org/10.1016/s1081-1206(10)60815-0

[23] Hashmi A, Ali W, Abulfaraj A, Binzagr F, Alkayal E. Enhancing Cancerous Gene Selection and Classification for High-Dimensional Microarray Data Using a Novel Hybrid Filter and Differential Evolutionary Feature Selection. Cancers. 2024; 16: 3913. https://doi.org/10.3390/cancers16233913

[24] Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Computers in Biology and Medicine. 2019; 112: 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

[25] Ejiyi CJ, Qin Z, Ukwuoma CC, Nneji GU, Monday HN, Ejiyi MB, et al. Comparative performance analysis of Boruta, SHAP, and Borutashap for disease diagnosis: A study with multiple machine learning algorithms. Network (Bristol, England). 2025; 36: 507–544. https://doi.org/10.1080/0954898X.2024.2331506

[26] Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. Statistics in Medicine. 1998; 17: 59–68. https://doi.org/10.1002/(sici)1097-0258(19980115)17:1%3C59::aid-sim733%3E3.0.co;2-7

[27] Ramteke M, Raut S. Prediction of polycystic ovary syndrome using machine learning with SFS and Boruta feature selection: an explainable AI approach. Systems Biology in Reproductive Medicine. 2025; 71: 439–460. https://doi.org/10.1080/19396368.2025.2560839

[28] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27: 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

[29] Lyu X, Zhang D, Pan H, Zhu H, Chen S, Lu L. Machine learning models for differential diagnosis of Cushing's disease and ectopic ACTH secretion syndrome. Endocrine. 2023; 80: 639–646. https://doi.org/10.1007/s12020-023-03341-7

[30] Long Y, Ren J, Cheng F, Duan Y, Wang B, Sun Y, et al. Identifying gray matter alterations in Cushing's disease using machine learning: An interpretable approach. Medical Physics. 2024; 51: 5479–5491. https://doi.org/10.1002/mp.17032

[31] Langarizadeh M, Moghbeli F. Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. Acta Informatica Medica. 2016; 24: 364–369. https://doi.org/10.5455/aim.2016.24.364-369

[32] Golounina OO, Belaya Zh.E., Voronov KA, Solodovnikov AG, Rozhinskaya LY, Melnichenko GA, et al. Machine learning methods in differential diagnosis of ACTH-dependent hypercortisolism. Problems of Endocrinology. 2024; 70: 18–29. https://doi.org/10.14341/probl13342 (In Russian)

[33] Zoli M, Staartjes VE, Guaraldi F, Friso F, Rustici A, Asioli S, et al. Machine learning-based prediction of outcomes of the endoscopic endonasal approach in Cushing disease: Is the future coming? Neurosurgical Focus. 2020; 48: E5. https://doi.org/10.3171/2020.3.FOCUS2060

[34] Demir AN, Ayata D, Oz A, Sulu C, Kara Z, Sahin S, et al. Machine Learning May Be an Alternative to BIPSS in the Differential Diagnosis of ACTH-dependent Cushing Syndrome. The Journal of Clinical Endocrinology and Metabolism. 2025; 110: e412–e422. https://doi.org/10.1210/clinem/dgae180