Machine Learning Models for Classification of Cushing's Syndrome Using Retrospective Data

Senol Isci, Derya Sema Yaman Kalender, Firat Bayraktar, and Alper Yaman

***Abstract*—Accurate classification of Cushing's Syndrome (CS) plays a critical role in providing the early and correct diagnosis of CS that may facilitate treatment and improve patient outcomes. Diagnosis of CS is a complex process, which requires careful and concurrent interpretation of signs and symptoms, multiple biochemical test results, and findings of medical imaging by physicians with a high degree of specialty and knowledge to make correct judgments. In this article, we explore the state of the art machine learning algorithms to demonstrate their potential as a clinical decision support system to analyze and classify CS to facilitate the diagnosis, prognosis, and treatment of CS. Prominent algorithms are compared using nested cross-validation and various class comparison strategies including multiclass, one vs. all, and one vs. one binary classification. Our findings show that Random Forest (RF) algorithm is most suitable for the classification of CS. We demonstrate that the proposed approach can classify CS with an average accuracy of 92% and an average F1 score of 91.5%, depending on the class comparison strategy and selected features. RF-based one vs. all binary classification model achieves sensitivity of 97.6%, precision of 91.1%, and specificity of 87.1% to discriminate CS from non-CS on the test dataset. RF-based multiclass classification model achieves average per class sensitivity of 91.8%, average per class specificity of 97.1%, and average per class precision of 92.1% to classify different subtypes of CS on the test dataset. Clinical performance evaluation suggests that the developed models can help improve physicians' judgment in diagnosing CS.**

***Index Terms*—Classification, cushing's syndrome, decision support systems, machine learning, prediction, random forest.**

I. Introduction

CUSHING's Syndrome (CS) is a potentially lethal disorder caused by abnormally high levels of cortisol hormone, first described in 1912 by Harvey Cushing [1], [2]. The estimated incidence of CS is 0.2–5 per 1 million per year, and its prevalence is 39–79 per million in various populations. The median age is 41.4 years, and the female to male ratio is 4 to 1 [3], [4]. It may stem from prolonged intake of glucocorticoids-steroid hormones that are chemically similar to natural cortisol, such as anti-inflammatory medications prescribed for asthma, rheumatoid arthritis, lupus, and other inflammatory diseases. Such hormones may also be taken after an organ transplant to suppress the immune system and prevent organ rejection. There are also endogenic causes in which the body produces an excessive amount of cortisol by itself. Cushing Disease, a form of CS, is the most common cause of excess endogenous cortisol production by the adrenal glands. It is caused by a pituitary tumor (i.e. adenoma which is usually a benign tumor in glands) that secretes an excessive amount of adrenocorticotropic hormone (ACTH), which then signals the adrenal glands to produce cortisol. CS might be a result of an adrenal gland tumor or adrenal hyperplasia (i.e. a genetic disorder in the adrenal gland), which can cause the adrenal gland to overproduce cortisol. Ectopic CS is another form of CS in which a tumor in another part of the body such as the pancreas, lung, or thyroid can result in CS by producing ACTH. It is called ectopic ACTH production because it is produced somewhere other than the pituitary gland.

Early diagnosis plays a crucial role in reducing mortality and improving the prognosis of this syndrome. However, the diagnosis of CS can be difficult, for instance, due to the gradual development of symptoms, and due to overlap with features of metabolic syndrome like increased blood pressure, high blood sugar, excess body fat around the waist, and abnormal cholesterol or triglyceride levels. Moreover, many of these features are common in the general population [5].

Laboratory investigations for patients with clinically suspected CS are divided into two stages. Stage-1 tests are screening tests for diagnostic purposes and applied to prove the presence of hypercortisolism. Stage-2 includes follow-up tests to evaluate the cause of hypercortisolism [4], [6], [7]. The most commonly used evaluation procedure for the exclusion or confirmation of CS has urine cortisol test which measures the cortisol level in a 24-hour sample of urine, saliva cortisol test which measures the cortisol level in the saliva, and low-dose dexamethasone test which measures the cortisol level in the blood after intake of

Manuscript received July 24, 2020; revised November 25, 2020 and December 23, 2020; accepted January 21, 2021. Date of publication January 29, 2021; date of current version August 5, 2021. *(Corresponding author: Alper Yaman.)*

Senol Isci is with the TUBITAK BILGEM Informatics, and Information Security Research Center, 41470 Kocaeli, Turkey (e-mail: senol.isci@tubitak.gov.tr).

Derya Sema Yaman Kalender is with the Department of Endocrinology, Faculty of Medicine, Izmir Katip Celebi University, 35620 Izmir, Turkey (e-mail: derya.sema@gmail.com).

Firat Bayraktar is with the Division of Endocrinology and Metabolism, Department of Internal Medicine, Dokuz Eylul University Medical School, 35340 Izmir, Turkey (e-mail: firat.bayraktar@gmail.com).

Alper Yaman is with the Department of Biomechatronic Systems, Fraunhofer Institute for Manufacturing Engineering, and Automation IPA Nobelstr. 12, 70569 Stuttgart, Germany (e-mail: alper.yaman@ipa.fraunhofer.de).

Digital Object Identifier 10.1109/JBHI.2021.3054592

the drug Dexamethasone. Despite the predictive value of these methods, cases with inconclusive test results still happen. Inconclusive results can be seen in patients with initial stages of this disease or in periodic forms of CS. Sometimes the diagnosis can only be made after long-term follow up or prolonged procedures, and may require hospitalization of the patient [6]. For patients with incidental adrenal mass, it is difficult to make a diagnosis or operation decision only with test results [8]. Incidental adrenal mass refers to the incidentally discovered adrenal mass during imaging which was not performed for suspected adrenal disease. Some conditions (e.g., obesity, diabetes mellitus, or depression) which have common features with CS may cause physiological hypercortisolism, and lead to incorrect results of the Dexamethasone suppression test (DST) [9].

Factors like laboratory errors, patient-induced errors, differences between groups, age, and gender may cause inconsistent test results. Traditionally, the diagnostic values of the test results have been analyzed using statistical methods and based on different selected cut-offs for sensitivity and specificity tradeoff. Therefore, the evaluation of the tests performed for the diagnosis and the identification of the cause may significantly vary. In many studies, different cut-off values have been found depending on the varying settings of the medical test and the employed statistical methods [10]. The diagnosis of CS and identification of its cause require careful and concurrent interpretation of signs and symptoms, results of multiple biochemical test measurements, and findings of medical imaging by physicians with a high degree of specialty and knowledge to make correct judgments.

We believe that these difficulties would be handled by utilizing a generalizable machine learning (ML) approach. In this article, we explore the state of the art ML algorithms, and demonstrate their usefulness as a clinical decision support system to evaluate results of the medical tests, and predict CS to facilitate the diagnosis and prognosis of CS. Furthermore, clinical performance evaluation of the suggested method was performed by comparing model predictions to the judgments of expert physicians.

We developed a publicly accessible web application (at https://cushings-syndrome-prediction.herokuapp.com/) as a clinical decision support tool for public use where a user can input the patient’ s test findings and receive prediction for CS. We used the Python scikit-learn ML library v0.21.3 [11], which contains implementations of all models used in this study. The software code is available at https://github.com/SenolIsci/cushing01.

A. Related Work

ML approach has been applied to CS related problems [12]. The automated interpretation of urine steroid profiles to classify normal and abnormal profiles of several metabolic conditions including CS has been studied [13]. Classification of CS using gene expression data of tumor tissues has been demonstrated [14]. The use of ML to identify predictors of early post-surgical and long-term outcomes in patients treated for Cushing disease (CD) has been studied [15]. Another study has aimed to identify facial anomalies associated with endocrinal disorders including CS using ML approach to facilitate the process of diagnosis and follow-up [16]. To the best of our knowledge, our study is the first to present a comprehensive investigation of the application of ML approach to diagnose CS and classify its subtypes using retrospective medical data.

II. Materials

A. Subjects

The retrospective medical records of 241 subjects (183 female and 58 male, age mean±standard deviation = 52.02±13.33 years) were used. The subjects were admitted to the endocrinology outpatient clinic of Dokuz Eylul University Medical Faculty due to CS symptoms or incidental adrenal adenomas between 2005 and 2016. We excluded 3 subjects of ectopic CS since data samples were not sufficient for further processing. The results of diagnostic tests including basal cortisol, basal ACTH, 1 mg DST cortisol, 2 mg DST cortisol, 8 mg DST cortisol, midnight cortisol, 24-hour urine cortisol (hospital reference range: 58–403 μg/24 h), and adrenal and pituitary imaging tests were examined. Dokuz Eylul University Faculty of Medicine Ethics Committee approved this research study and waived the requirement for consent (2019/28-26).

B. Dataset Profile

The CS dataset consists of 241 samples and 11 features (i.e. predictor variables). The target variable represents 3 classes of subtypes (i.e., pituitary (PT), adrenal (AD), subclinical (SC)) for patients with CS and 1 class (i.e., nonfunctional adrenal adenoma (NF)) for patients without CS. The statistical properties (i.e. mean, standard deviation, ratio, and counts (n)) of the dataset features according to the type of diagnosis are given in Table I. Dataset has missing values because some of the medical tests were not performed depending on the symptoms and results of the accompanying tests as decided by the physicians. A number of the feature distributions were observed to be skewed (i.e. asymmetry of the distribution about its mean), especially all the distributions of cortisol and ACTH related features. Dataset faces a moderate class imbalance problem because one of the classes (i.e. NF) is represented by a large number of samples whereas the others are represented by fewer samples. Age and gender features were reported for reference purposes, and they were not used in the rest of the study.

III. Technical Approach

Fig. 1 shows a flowchart of the technical approach. The following sections provide the details of steps, procedures, and algorithms of the overall technical approach.

A. Dataset Splitting and Subset Preparation

The input dataset was randomly split into T&V (Train and Validation) dataset (n = 168, 70%) and FIT dataset (n = 73, 30%) by stratified sampling ensuring that ratios of classes are represented in the newly created datasets. Due to the limited dataset size, two-sample Kolmogorov-Smirnov (K-S) test was performed for each feature to check whether both T&V and FIT datasets are representative of the same distribution. The procedure is tested for repetitiveness over 10 sampling cases

TABLE I
PROPERTIES OF THE DATASET

			Diagnosis			
Feature	Description	Missing Value Ratio (%)	NF (n=104)	SC (n=59)	AD (n=42)	PT (n=36)
age	subject age (years)	0%	55.24±10.46	56.88±12.65	47.90±13.7	39.56±12.69
gender	subject gender (female/male count)	0%	71/33	47/12	33/9	32/4
bc	basal cortisol (μg/dL)	2%	13.84±6.05	13.55±4.48	20.64±7.82	21.38±10.90
bacth	basal ACTH (pg/mL)	2%	20.88±14.15	10.62±7.00	9.09±8.59	59.08±46.89
1mgDSTc	cortisol after 1mg DST (μg/dL)	11%	1.37±1.24	4.63±4.22	13.30±8.43	12.48±5.70
2mgDSTc	cortisol after 2mg DST (μg/dL)	49%	1.77±1.02	4.36±3.47	14.58±10.69	12.30±7.33
8mgDSTc	cortisol after 8mg DST (μg/dL)	71%	-	5.00±5.61	14.79±10.99	5.15±5.59
mc	midnight cortisol (μg/dL)	32%	3.87±2.63	6.43±3.33	16.16±6.71	18.69±9.61
ufc	urinary free cortisol (μg/24h)	33%	156.32±79.63	181.29±157.71	307.04±294.07	339.34±286.12
adrMass	mass in adrenal imaging (yes/no count)	0%	104/0	59/0	42/0	8/28
pitMass	mass in pituitary imaging (yes/no count)	0%	1/103	8/51	4/38	36/0

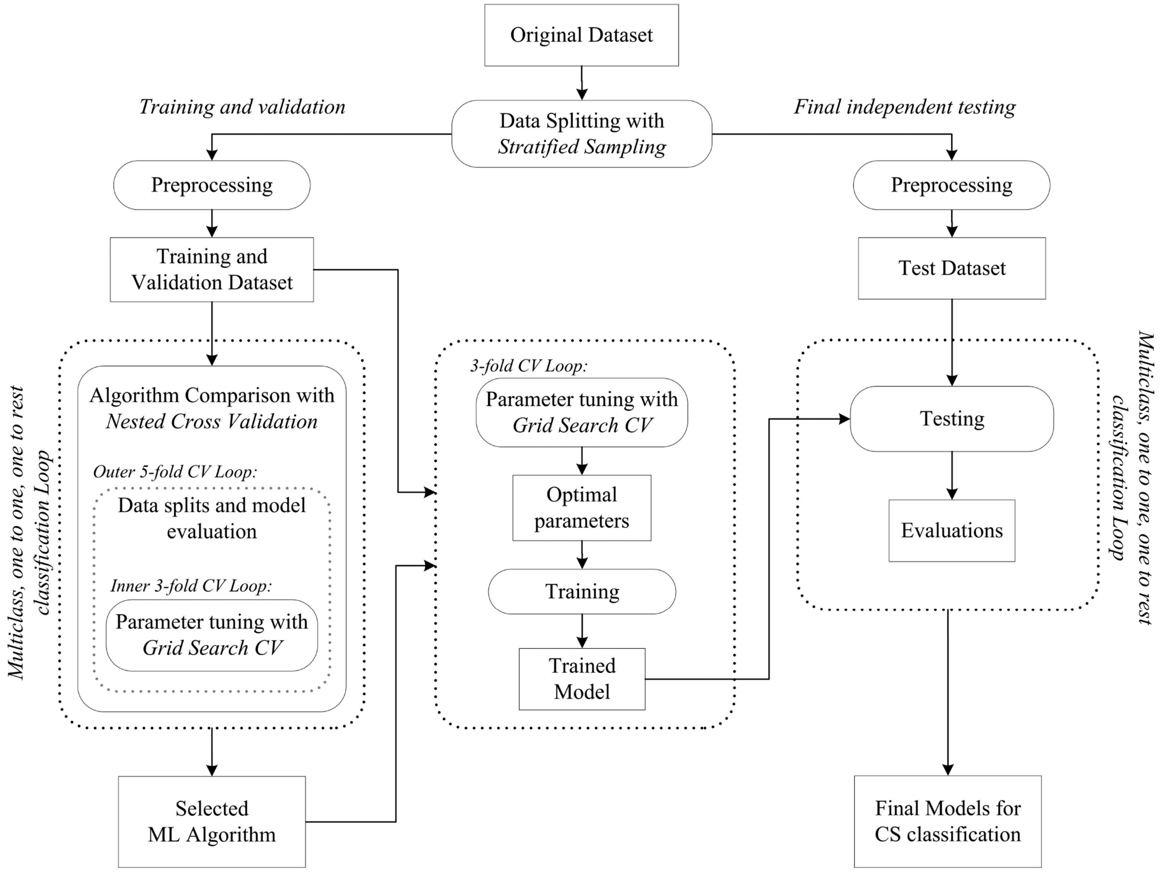
Original Dataset
Data Splitting with Stratified Sampling
Training and validation
Final independent testing
Preprocessing
Preprocessing
Training and Validation Dataset
Test Dataset
Algorithm Comparison with Nested Cross Validation
Outer 5-fold CV Loop:
Data splits and model evaluation
Inner 3-fold CV Loop:
Parameter tuning with Grid Search CV
Multiclass, one to one, one to rest classification Loop
3-fold CV Loop:
Parameter tuning with Grid Search CV
Optimal parameters
Training
Trained Model
Testing
Evaluations
Multiclass, one to one, one to rest classification Loop
Selected ML Algorithm
Final Models for CS classification

Fig. 1. The flowchart of the overall method.

and randomly chosen data subsets were used. Two-sample K-S test with 5% significance level yielded high values of p-values between 0.354 and 1 and K-S statistics between 0.003 and 0.128, indicating that the datasets were representative of the same distribution.

We intend to evaluate multiclass, one vs. all, and one vs. one binary classification strategies for the task of CS type classification. Therefore, the data subsets were created accordingly for these class comparison strategies (n = 11). Multiclass strategy consists of fitting one classifier for all classes, namely, PT, AD, SC, and NF. One vs. one strategy consists of fitting one classifier per class pair. One vs. all strategy consists of fitting one classifier per class, and for each classifier, the class is fitted against all the other classes. The dataset properties according to class comparison strategy are listed in Table II. ALL refers to all classes in multiclass setting whereas it refers to the rest of the classes in binary setting. This means that class labels were merged together for classes in the same comparison. For Stage-1, bc, bacth, 1 mgDSTc, mc, and ufc features were used. For Stage-2, in addition to the features used in Stage-1, 2 mgDSTc, 8 mgDSTc, adrMass, and pitMass features were used.

B. Preprocessing

1) Missing Value Imputation: Physicians may decide to skip some of the tests if the previous tests are sufficient for the diagnosis. For example, 8 mg DST test is performed for identifying the type of CS only after the screening tests such as 1 mg DST. If a subject is believed to have nonfunctional adrenal adenoma,

TABLE II
CLASS COMPARISON STRATEGIES, ASSOCIATED DATASETS, AND FEATURES

Class Comparison Type	**T&V dataset Number of observations in each class**	**FIT dataset Number of observations in each class**
Multiclass		
ALL	[73 41 29 25]	[31 18 13 11]
One vs. one binary		
PTvsAD	[29 25]	[13 11]
ADvsNF	[73 29]	[31 13]
PTvsNF	[73 25]	[31 11]
SCvsNF	[73 41]	[31 18]
ADvsSC	[41 29]	[18 13]
PTvsSC	[41 25]	[18 11]
One vs. all binary		
SCvsALL	[127 41]	[55 18]
PTvsALL	[143 25]	[62 11]
ADvsALL	[139 29]	[60 13]
ALLvsNF	[73 95]	[31 42]

8 mg DST is almost always skipped. Depending on the severity of symptoms and other indicators, the physician may decide to skip low dose DST test and jump to 8 mg DST test. Therefore, we avoided omitting features with missing values, and we used imputation.

The treatment of missing data is a broad statistical problem, and there is no universal imputation method performing best in every situation [17]. The simplest option is discarding samples with missing values. However, in this study, dropping samples with missing values of features was not considered as an option because missing values are ubiquities in the original dataset. This means that such omissions would result in severe information loss and a very small dataset. The missing value for a given feature was replaced by the median of all known values of that feature calculated separately both for T&V dataset and FIT dataset. In this way, information leakage is prevented, and independence of the FIT dataset is preserved from the rest of the training process, however, at the expense of degrading statistics.

2) Skewed Feature Distributions Problem: In general, skewed distributions of the feature in the dataset will degrade the model's ability to describe more prevalent cases to deal with much rarer cases that happen to take extreme values. Some ML algorithms (e.g. LDA classifier) have normality assumption for the underlying populations, and their performance is adversely affected by the violation of this assumption [18]. Classifiers that are free of any distributional assumptions (e.g. RF) are expected to perform well with a variety of distributions as long as the class distributions are reasonably distinct [19]. In this study, we applied logarithmic transformation (log base 10) to the dataset to make feature distributions less skewed and reported the level of skewness using Fisher-Pearson coefficient of skewness [20], [21]. Mean absolute skewness was reduced from 2.398 to 0.897 after the transformation of the T&V dataset.

C. Machine Learning Algorithms

We compared several ML algorithms, all of which can be employed in multiclass, one vs. all, and one vs. one binary classification strategy for the task of classification of CS. Multiclass strategy encompasses the fitting of one classifier for all classes. One vs. one strategy includes fitting one classifier per class pair. One vs. all strategy consists of fitting one classifier per class, assuming the label of the class as positive and labels of other classes as negative. The algorithms evaluated are Support Vector Machine (SVM) [22], K-nearest Neighbor (KNN) [23], Logistic Regression (LG) [24], Linear Discriminant Analysis (LDA) [25], Decision Tree (DT) [26] based on the Classification and Regression Tree (CART) algorithm, Random Forest (RF) [27], Adaptive Boosting (AdaBoost) [28], and Gradient Boosting (GB) [29]. Generalization of AdaBoost for multiclass classification has also been introduced, which is referred to as AdaBoost-SAMME [30]. In our study, we used CART DT as a weak learner in the AdaBoost-SAMME model.

D. Performance Metrics and Measurement Tools

We presented our results using basic evaluation metrics derived from confusion matrix associated with the classifier: the area under the Receiver Operating Characteristics (ROC) curves (ROC AUC), accuracy (ACC) measured by ROC AUC, the area under the curve (PRC AUC) values of the precision-recall curve (PRC) for the models, Sensitivity or Recall (SENS), Specificity (SPEC), Precision (PREC), and F1 Score. We also employed average per class F1 score ($\mathrm{F1_m}$) for multiclass or binary classification setting, which is calculated by averaging F1 scores over all classes. $\mathrm{F1_m}$ was used as the primary metric for comparisons throughout the study. The mean and variance of AUC were calculated over 5 ROC curves. The variance roughly shows how the classifier output is affected by changes in the training data. PRC shows the trade-off between precision and recall of different thresholds.

In the diagnosis of CS and identification of its cause, one can interpret the outcome of the model and determine the most impactful medical tests by "feature importance". Therefore we present relative importance of features in the selected models using the mean decrease in impurity calculation [31].

E. Definition, Comparison, and Selection of the Best Algorithm

We aim to evaluate ML algorithms in terms of their predictive performance (i.e. generalization accuracy) on unseen independent data and identify the ML algorithm that is best suited for the diagnosis of CS. We compared trained models from the algorithm's model space and selected the best performing models by tweaking the hyperparameters of the algorithm. Furthermore, we aim to optimize hyperparameters from a given hyperparameter space by comparing and selecting values of hyperparameters with respect to their performance in minimizing error. Nested cross-validation (NCV) approach is well suitable for these tasks under limited data size and produces almost unbiased performance estimates [32]. NCV is relatively straightforward as it is a nesting of two k-fold cross-validation (CV) loops: the inner loop is responsible for the model selection and hyperparameter optimization, and the outer loop is responsible for estimation of the generalization accuracy for model evaluation.

The classifiers for the class comparisons were listed in Table II. Class comparisons were iterated for the features of diagnostic Stage-1 and diagnostic Stage-2. In the end, 22 classifier models were evaluated per each algorithm.

We evaluated each algorithm and its associated models with a 5 × 3 NCV procedure with steps as the following (Fig. 1): the input T&V dataset was divided into stratified 5 folds (i.e. each fold contains approximately the same percentage of samples of each target class as in the input dataset.). For each iteration in the outer loop, 1 data fold was reserved as held-out validation data for model evaluation. The remaining 4 folds were passed to the inner loop where model parameter tuning was performed. All the models have hyperparameters that must be compared and selected. These include, for example, penalty term for overfitting in LG, the kernel coefficient in SVM, and the number and depth of trees in RF. The input data passed from the outer fold was divided into stratified 3 folds. The inner loop included a grid search over candidate hyperparameters using 2 folds of data. Each parameter setting was evaluated with the remaining 1 fold validation data. The inner loop was repeated 3 times, each time holding out a different validation fold. The hyperparameters that yielded the best average CV score were selected and reported back to the outer loop. The model was then trained on the data in the 4 folds using the best parameters passed from the inner loop and then evaluated for its predictive performance using the 1 fold held-out validation data in the outer loop. This process was repeated 5 times in the outer loop, resulting in evaluations of model performance 5 times, and the average evaluation score was obtained from averaging outer held-out validation datasets. Scores using several evaluation metrics were reported. NCV procedure was repeated for each of the classification algorithms (n = 8) and each class comparison strategy (n = 11) using a different feature set for each of the diagnostic stages (n = 2). Overall, NCV was performed 8 × 11 × 2 = 176 times. The best performing algorithm was selected according to the highest grand average for all runs.

F. Training, Testing, and Evaluation of the Best Algorithm

Following the algorithm selection, the selected algorithm was used to build classification models with different class comparison strategies and feature sets which were selected after discussing it with physicians specialized in CS as to how useful it will be in clinical diagnosis and prognosis. We created similar classifiers as listed in Table II for multiclass comparison for all class types, one vs. one binary comparisons, and one vs. all binary comparisons. Class comparisons are iterated for the features of diagnostic Stage-1 and diagnostic Stage-2.

The hyperparameters used in the models were determined by 3-fold CV hyperparameter search approach similar to the one in the algorithm selection procedure. However, this time, all of the T&V dataset was used in the search.

In the training phase, ML model was inferred from T&V dataset with known class labels. The parameters of the model were optimized by fitting the observations in the dataset to the output target variable. Afterward, the trained models were evaluated on the FIT dataset for their predictive performance (i.e. generalization accuracy) for unseen data. At the final step, the complete dataset was used to train final classification models to be used for testing new samples.

TABLE III
OVERALL RESULTS OF ALGORITHM COMPARISON

	Stage-1 Scores		Stage-2 Scores		Overall Average Scores	
Algorithm	**F1$_{\text{m}}$**	**ACC**	**F1$_{\text{m}}$**	**ACC**	**F1$_{\text{m}}$**	**ACC**
RF	**88.1±9**	**90.3±8**	**90.9±8**	**92.5±7**	**89.5±9**	**91.4±8**
SVC	85.5±11	87.7±10	88.7±9	90.6±7	87.1±10	89.2±9
LG	85.4±13	87.5±11	88.2±10	90.2±8	86.8±11	88.8±10
AdaBoost	85.5±10	87.9±9	87.3±11	89.4±10	86.4±11	88.6±10
LDA	83.7±15	87.9±9	88.5±12	90.8±9	86.1±14	89.4±9
KNN	84.3±12	87.2±10	87.7±11	90.2±9	86.0±11	88.7±10
DT	84.7±11	87.2±10	87.2±11	89.1±10	86.0±11	88.2±10
GB	82.2±11	85.6±10	85.0±12	87.6±11	83.6±12	86.6±10

G. Clinical Performance Evaluation

It is of importance to assess calibration to see how reliable is the predicted outcomes of the models for clinical usefulness. Calibration refers to the level of agreement between observed outcomes and predictions [33]. We present the calibration curve for the model.

The clinical performance of ML models was evaluated by comparing their predictions to the judgments made by human experts. For this comparison, patient test dataset of 73 cases (i.e. FIT dataset) was used. A group of 4 physicians specialized in CS was asked separately to make judgments about whether a patient has CS or not by assigning probability values to each class label using Stage-1 and Stage-2 features through all samples in the test dataset. We, then, applied soft voting to combine expert predictions. Namely, the predicted class probabilities provided by each expert were collected and averaged. The final class label was derived from the class label with the highest average probability.

IV. RESULTS

A. Algorithm Comparison

Table III summarizes the algorithm comparison results in terms of performance estimates in classification of each of 4 classes (i.e. NF, SC, AD, and PT) computed by 5 × 3 NCV procedure on the T&V dataset. It also includes overall average scores over all models (n = 22). Each score in Stage-1 and Stage-2 columns in Table III is the mean of performance scores averaged over validation folds in 5 × 3 NCV for each model (n = 11) created for all class comparison strategies.

The highest F1$_{\text{m}}$ scores for Stage-1 (88.1±9), Stage-2 (90.9±8), and overall average (89.5±9) were achieved by RF. Other algorithms had comparable average scores to RF algorithm, but RF models had smaller standard deviations. It could be seen from these results that models built using RF showed better performance compared to the other algorithms for both Stage-1 and Stage-2 diagnosis. It was also observed that model performance increases with the increase in the number of features. This is an indication that important features were added to the model building and that promoted the prediction power.

Table IV gives the decomposition of NCV scores of ML algorithms averaged over Stage-1 and Stage-2 against class comparison strategies. In 6 of 11 class comparison strategies, RF ranked first. The highest score for RF was 97.8±7 achieved

TABLE IV
Comparison of Algorithms Against Class Comparison Strategy

	ALL	**ALLvsNF**	**SCvsALL**	**PTvsALL**	**ADvsALL**	**PTvsAD**	**ADvsNF**	**PTvsNF**	**SCvsNF**	**ADvsSC**	**PTvsSC**
Algorithm						**$F1_m$ %**					
RF	**81.3±4.0**	**91.7±4.0**	**81.4±8.0**	93.3±8.0	**86.4±7.0**	93.1±12.0	93.2±7.0	**97.8±7.0**	90.8±8.0	**82.1±9.0**	93.5±5.0
SVC	78.3±5.0	89.6±5.0	74.6±11.0	92.1±6.0	83±8.0	95.0±10.0	92.6±7.0	95.1±7.0	85.5±7.0	80.8±8.0	91.9±8.0
LG	**81.3±4.0**	88.1±4.0	67.4±9.0	94.4±6.0	79.6±9.0	95.9±10.0	**94.1±6.0**	96.7±6.0	85.1±9.0	78.5±8.0	**93.6±8.0**
AdaBoost	70.9±8.0	90.5±3.0	80.0±6.0	91.5±8.0	85.4±6.0	90.0±14.0	92.7±7.0	93.7±7.0	90.3±5.0	80.5±9.0	84.6±15.0
LDA	80.6±5.0	89.5±4.0	59.0±14.0	**96.1±5.0**	77.6±14.0	**97.1±7.0**	93.4±6.0	97.2±7.0	84.8±9.0	78.9±10.0	92.6±7.0
KNN	75.8±5.0	89.6±5.0	73.4±9.0	92.7±9.0	78.6±8.0	96.2±7.0	91.3±8.0	97.2±7.0	81.4±9.0	76.3±11.0	93.5±8.0
DT	75.9±9.0	86.6±6.0	74.0±10.0	89.7±9.0	84.1±7.0	90.1±13.0	90.3±9.0	94.3±8.0	**91.9±7.0**	77.8±15.0	91.1±4.0
GB	74.6±6.0	88.2±3.0	71.1±11.0	89.1±9.0	78.6±10.0	89.2±12.0	88.9±10.0	94.2±8.0	83.6±8.0	71.8±15.0	90.4±5.0

TABLE V
Best Hyperparameters for RF Models Obtained From Cross-Validated Grid Search

Class Comparison Strategy	Feature Set	Number of Trees	Class Weight	Split Criterion	Max Depth of Trees	Mean $F1_m$%
ALL	Stage-1	50	balanced subsample	gini	5	77.3
ALL	Stage-2	100	balanced	gini	6	84.1
ALLvsNF	Stage-1	100	balanced	entropy	6	90.7
SCvsALL	Stage-2	100	balanced subsample	gini	4	84.8
ADvsALL	Stage-2	100	balanced	entropy	4	88.1
PTvsALL	Stage-2	100	balanced	gini	4	98.8

in PTvsNF comparison whereas the lowest score was 81.3±4 for multiclass comparison ALL. It was seen that SCvsALL, ADvsSC, and multiclass ALL classifications were the cases in which algorithms generally achieved lower performance.

According to NCV results, RF algorithm was found to be the best performing algorithm and selected for in-depth study. A subset of RF models from 11 different class comparison strategies was selected for training after discussing the clinical methods and models with physicians specialized in CS. Usually, at Stage-1, it is aimed to diagnose patients with hypercortisolism. Discrimination between nonfunctional adrenal adenoma and CS regardless of its subtype is required. Therefore, ALLvsNF one vs. all binary classification model with Stage-1 features was employed. At Stage-2, identification of a specific CS subtype against the rest of alternative CS subtypes is required. Therefore, SCvsALL, ADvsALL, and PTvsALL one vs. all binary models with Stage-2 features were employed. Multiclass classification was employed with both Stage-1 and Stage-2 features.

B. Parameter Optimization of the Best Algorithm

Table V shows the optimized hyperparameters and mean performance scores obtained from 3-fold cross-validated exhaustive search over hyperparameter values for selected RF models using the complete T&V dataset. Optimized hyperparameters were found to be the number of decision trees employed in the RF, max depth of decision trees, split criterion, and class weight criterion.

C. Training of the Best Algorithm

Fig. 2 shows the cross-validated learning curves of ALLvsNF model at Stage-1 and Multiclass ALL model at Stage-2 in training phase. CV scheme is employed due to the limited dataset size and selected to be based on stratified random sampling with replacement in 20 random splits with 80% training and 20% validation sets with preserved class proportions. Afterwards, the scores are averaged over all 20 runs for each training subset size and plotted against the varying data size. For multiclass ALL model at Stage-2, high scores in the learning curve starting from the low data sizes indicate low bias, and this is usually typical for tree-based algorithms. Training and validation curves appear to be converging but there is slight variance (i.e., the gap between curves) in both models. These characteristics show slight overfitting in the learning procedure. Overfitting may be reduced and

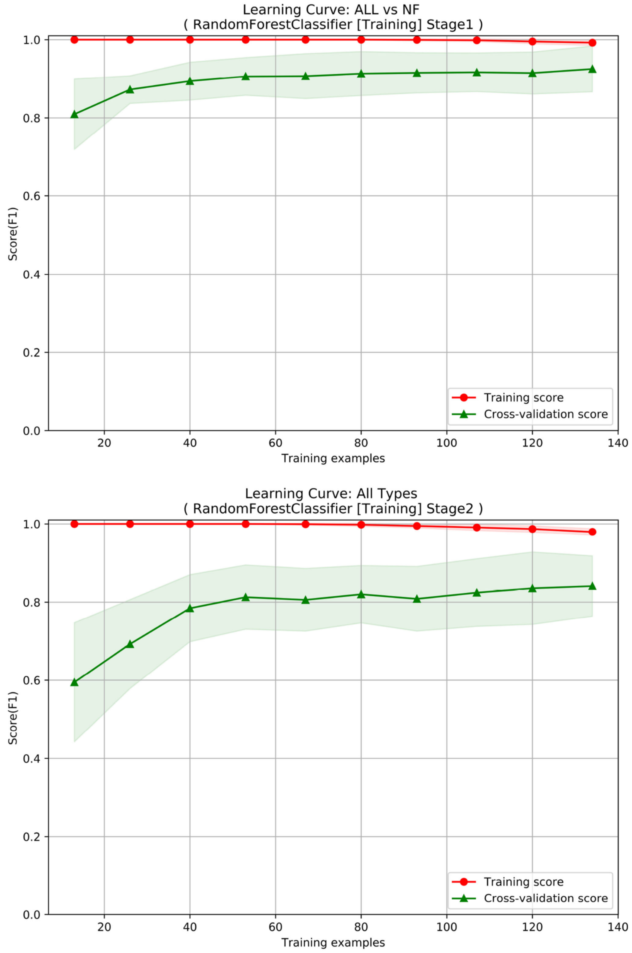

Fig. 2. Cross-validated learning curves of ALLvsNF model at Stage-1 (top) and multiclass ALL model at Stage-2 (bottom) in training phase.

TABLE VI
TRAINING RESULTS OF RF MODELS

Classes	Feature Set	$F1_m$%	ACC%	F1%	SENS%		SPEC%		PREC%		ROC AUC%	PRC AUC%
ALL	Stage-1	95.0	95.8	-	NF	97.3	NF	100.0	NF	100.0	-	-
					SC	92.7	SC	97.6	SC	92.7		
					AD	96.6	AD	97.1	AD	87.5		
					PT	96.0	PT	100.0	PT	100.0		
ALL	Stage-2	96.4	97.0	-	NF	95.9	NF	100.0	NF	100.0	-	-
					SC	95.1	SC	98.4	SC	95.1		
					AD	100.0	AD	97.8	AD	90.6		
					PT	100.0	PT	100.0	PT	100.0		
ALLvsNF	Stage-1	98.5	98.2	97.9	100.0		95.9		96.9		99.8	99.9
SCvsALL	Stage-2	91.3	94.6	95.6	97.6		93.7		83.3		98.3	95.9
ADvsALL	Stage-2	90.3	95.8	97.4	100.0		95.0		80.6		99.5	98.2
PTvsALL	Stage-2	100.0	100.0	100.0	100.0		100.0		100.0		100.0	100.0

generalization accuracy may be improved by increasing the size of the dataset as more samples become available.

Table VI shows the performance results of the trained RF models. High sensitivity and specificity were achieved in all models, ranging from 92.7% to 100% and from 93.7% to 100%, respectively. ALLvsNF model had 100% sensitivity and was able to correctly catch all CS samples at Stage-1. It also correctly generated a negative result for NF samples at 95.9% specificity. Precision or Positive Predictive Power (PPV) of ALLvsNF model was 96.9% that means only 3.1% of positive results generated by the model were actually NF.

PTvsALL achieved the highest classification score ($F1_m$) of 100% whereas ADvsALL had the lowest score of 90.3%. AD type had the lowest precision of 87.5% compared to other types in the multiclass ALL model. Multiclass ALL model at Stage-2 achieved 96.4% classification score ($F1_m$) and improved on sensitivity and specificity compared to the multiclass ALL model at Stage-1.

All ROC curves have high mean AUC values: 0.9690±0.017 for ALLvsNF, 0.947±0.074 for ADvsALL, 0.926±0.022 for SCvsALL, and 0.995±0.000 for PTvsALL. This means that classifiers are better at classifying positive and negative observations. CV roughly shows how the classifier output is affected by changes in the training data. Standard deviations calculated for CV folds indicate that PTvsALL model is more robust to changes in the data whereas ADvsALL model is more susceptible to data perturbations.

D. Feature Importance

Fig. 3 shows the relative feature importance levels of the trained models. For multiclass ALL model at Stage-1, the features 1 mgDSTc, bacht, and mc were found to be the relatively most important features that contribute most to the classification performance. For ALLvsNF model, 1 mgDSTc was by far the most important feature and mc being the second. Relatively most important features for multiclass ALL model at Stage-2 were inferred to be 1 mgDSTc, bacth, pitMass, and mc. For SCvsALL model, 1 mgDSTc, 2 mgDSTc, bacth and mc rank at top. The features mc, 1 mgDSTc, and bacth were found to be the most important features for ADvsALL model, For PTvsALL model, pitMass was by far the most important feature, accompanied by bacth and 8 mgDSTc.

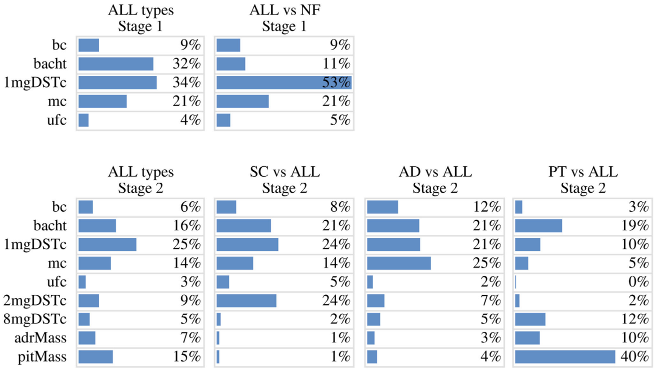

Fig. 3. Feature importance levels of RF models.

We highlight that imaging alone is not adequate to correctly identify and classify CS. For example, 40% of patients with proven Pituitary CS have normal pituitary MRI [7]. The sensitivity of computed tomography (CT) is between 40–50%, whereas the sensitivity of magnetic resonance imaging (MRI) is in the range of 50–60%. This low sensitivity is due to the average size of corticotropic adenomas (i.e., benign tumor in the corticotropic cells of the pituitary gland) being 5–6 mm. Some of these tumors are 1–3 mm. Ectopic pituitary adenomas (i.e., tumor of pituitary tissue found in sinus or nasal cavity) are one of the false-negative causes in diagnosis based on MRI. In people aged 30–40 years, incidental pituitary adenomas occur around 10%. That is, no imaging results should be interpreted without biochemical results [4].

This situation is also reflected in our dataset. For some samples, it is seen that imaging results overlap. For example, 8 patients out of 36 patients with proven Pituitary CS have positive pituitary imaging but they also have positive adrenal imaging results. Four patients out of 42 patients with proven Adrenal CS have positive adrenal imaging results as well as positive pituitary imaging result. Eight patients out of 59 Subclinical CS patients have both positive adrenal and pituitary imaging results. It is also noted that imaging adds to clinical value. The overall F1 score achieved by features including adrenal and pituitary imaging was 92.1% as reported in Table VII. To elaborate, we ran the model without imaging features, and overall F1 score was reduced to 85.7%.

TABLE VII
TEST RESULTS OF RF MODELS

Classes	Feature Set	$F1_m$%	ACC%	F1%	SENS%	SPEC%	PREC%	ROC AUC%	PRC AUC%
ALL	Stage-1	84.1	84.9	-	NF 90.3 SC 72.2 AD 76.9 PT 100.0	NF 91.9 SC 92.5 AD 98.1 PT 94.4	NF 90.3 SC 76.5 AD 90.9 PT 78.6	-	-
ALL	Stage-2	92.1	91.8	-	NF 93.5 SC 88.9 AD 84.6 PT 100.0	NF 97.4 SC 92.7 AD 100.0 PT 98.2	NF 96.7 SC 80.0 AD 100.0 PT 91.7	-	-
ALLvsNF	Stage-1	93.8	93.2	92.1	97.6	87.1	91.1	96.9	97.2
SCvsALL	Stage-2	81.6	86.3	80.5	72.2	90.9	72.2	93.8	71.6
ADvsALL	Stage-2	97.6	95.9	87.0	76.9	100.0	100.0	94.9	90.6
PTvsALL	Stage-2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

E. Final Testing

Table VII shows the results of the testing of trained models using FIT dataset to show their generalization accuracy to unseen independent samples. ALLvsNF model at Stage-1 achieved the highest $F1_m$ score (93.8%), ROC AUC (0.969), and PRC AUC (0.972) with 97.6% sensitivity, 87.1% specificity, and 91.1% precision.

For Stage-1, the sensitivity value for PT class was 100%, which means that multiclass ALL model was able to catch all PT samples. It had moderate sensitivity values for SC and AD, respectively. All specificity values for classes were high, ranging from 91.9% to 98.1%. Precision values for NF and AD were above 90%. However, precision values were moderate for SC and PT. Multiclass ALL model at Stage-2 improved in all aspects compared to the Stage-1 multiclass ALL model.

SCvsALL model had 81.6% classification performance score ($F1_m$), a moderate sensitivity of 72.2%, high specificity of 90.9%, and moderate precision of 72.2%. It is also seen ADvsALL model had high classification performance ($F1_m$) of 97.6%, moderate sensitivity of 76.9%, maximum specificity of 100%, and maximum precision of 100%. PTvsALL model had maximum score, sensitivity, specificity, and precision of 100%. The model was able to correctly discriminate all PT type samples and samples of the rest of the classes.

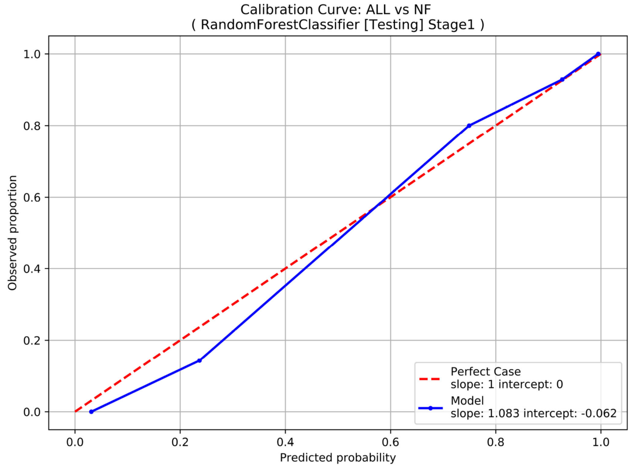

Fig. 4. Calibration curve for the ALL vs NF Model at Stage1.

TABLE VIII
CLINICAL PERFORMANCE COMPARISON OF ML MODELS VS. HUMAN EXPERTS

Predictor	**Classes**	**Feature Set**	**$F1_m$%**	**ACC%**
ML	ALLvsNF	Stage-1	93.8	93.2
Human Expert	ALLvsNF	Stage-1	84.6	86.3
ML	ALL	Stage-2	92.1	91.8
Human Expert	ALL	Stage-2	72.2	72.6

F. Clinical Performance Evaluation

The calibration curve for the ALL vs NF Model at Stage1 is shown in Fig. 4. Predictions (i.e., mean of binned predicted probabilities) are on the x-axis. The observed proportions (i.e., proportion of samples whose class is the positive class) are on the y-axis. Perfect case is represented by 45° line with a slope of 1 and intercept of 0 on the x-axis. The calibration curve of the model has slope of 1.083 and intercept of -0.062, and is close to 45° line. This supports that predictions made by the model are reliable.

The clinical performances of ML models and human experts were listed in Table VIII. In Stage-1, ML models achieved better predictions with $F1_m$ score of 93.8% compared to human experts with $F1_m$ score of 84.6%. In Stage-2, human experts achieved lower scores. The results reveal that ML model is able to find underlying patterns in the CS data with a constraint set of features among biochemical tests, presence of other illnesses, or other clinical findings.

V. DISCUSSIONS

ML algorithm comparison results indicate that most of the compared algorithms achieve good estimates of predictive performances. However, RF algorithm is found to be the best algorithm in overall performance according to the scores achieved for different class comparison schemes and feature sets. The algorithms generally show better performance with features of Stage-2 than with features of Stage-1. This is probably due to the inclusion of more informative features into the models, and it helps better fitting of model parameters to the data. This is actually observable in the relative feature importance levels of RF-based multiclass models that were inferred in the training phase as seen in Fig. 3. For the RF-based multiclass model at Stage-1, bacth, 1 mgDSTc, and mc contributed much to the predictive power so that the model achieved classification accuracy of 95% (See Table VI). Similarly, for the RF multiclass model at Stage-2, the same features were found to be the most important features besides the newly added pitMass feature, and the model achieved an improved classification accuracy of 96.4%.

Although models based on one vs. one class comparison strategy technically show good performance results at the algorithm selection step, one vs. one binary models were excluded from further training and testing because of their limited clinical use. For example, for ADvsSC binary model, one has to eliminate other classes beforehand, and a dataset consisting of only these two classes is required for the classification results to have real meaning. This looks infeasible because one has to first eliminate NF and PT classes from the samples, which are not yet known. Moreover, the biochemical tests performed at Stage-1 are generally aimed to discriminate nonfunctional adrenal adenoma and CS regardless of its subtype. Clinical testing at Stage-2 aims to discriminate the cause of CS such as adrenal CS, ectopic CS, or pituitary CS.

Ectopic CS samples were not utilized in model training due to the lack of enough samples. Therefore, our system cannot directly classify ectopic CS. Approximately 20% of ACTH-dependent cases are ectopic CS whereas approximately 80% of ACTH-dependent CS is pituitary CS (i.e. Cushing Disease) [4], [7]. Since both pituitary CS and ectopic CS are ACTH-dependent, some of the samples predicted as pituitary CS in the original dataset may actually be ectopic CS. Therefore, further study needs to be done by the physicians to diagnose correctly. Bilateral inferior petrosal sinus sampling (BIPSS) test, imaging findings, bacth levels, and rate of suppression in cortisol level after 8 mg DST are usually informative in separating ectopic CS from pituitary CS. As we gather more data samples for ectopic CS, the models can be easily updated to discriminate ectopic CS as well.

Median imputation induces a bias in the relationship between features and the target variable, and may not perform as good as more elaborate methods such as KNN, Maximum Likelihood, and Multivariate imputation by chained equations (MICE). However, such approaches introduce additional parameters, and this may lead to errors due to unsuccessful tuning of the parameters, and eventually reduced generalizability in case of limited data size [34]–[36]. As we continue to collect more data, we intend to evaluate alternative imputation methods.

Class imbalance in the dataset is known to reduce the predictive performance of a model [37]. Several methods exist to alleviate this problem, such as downsampling and oversampling methods. Since the dataset size is relatively small, downsampling the majority class causes information loss. Oversampling is likely to introduce bias to the accuracy since the new data samples are generated from a few old samples, and they cannot introduce much variance to the dataset. Therefore, we address this problem by using ML algorithms that can inherently handle imbalanced data classification with class weighting in the learning process.

The ML approach outperforms the human expert judgments in clinical performance evaluation. This is mainly because of the expert' s failure in discriminating the cases with subclinical CS. More than 20 characteristic signs and symptoms have been reported for CS [12]. These and other features make the diagnosis complex and sometimes confusing. It is also known that subclinical CS does not have typical signs and symptoms of hypercortisolism. The mild cortisol secretion may cause hypertension, central obesity, impaired glucose tolerance or diabetes, hyperlipemia, and osteoporosis. However, these complications are frequent in the population and cannot be directly attributed to the subclinical CS [38]. As a consequence, physicians have to diagnose subclinical CS by checking the biochemical test results as well as the aforementioned metabolic complications after a follow-up period.

The interpretation of medical tests to diagnose CS and classify its subtypes is time-consuming and limited by the physicians' capacity and experience to integrate numerous and complex information. Results from studies in other hospitals and medical centers may vary, and be related to factors such as laboratory errors, patient induced errors, differences between groups, age, and gender. We demonstrated that an ML-based decision support system might help.

Our approach is adaptable to new data and will improve as new samples are gathered. Once trained, the prediction models require very low computational resources. Furthermore, the features of the models are derived from tests routinely collected in the hospital. It can also serve as a general framework and allows the integration of data from different hospitals and medical centers. These models can help to screen a large portion of negative cases at the early stages of clinical diagnosis, prognosis, and treatment.

VI. Conclusion

We compared several prominent ML algorithms and demonstrated the ability of the RF-based models to accurately predict clinical interpretation of CS, despite the moderate size of the dataset and class imbalance problem. We think that the success of the RF algorithm is because of its capabilities of handling small sample and imbalanced CS data, learning complex dependencies, reducing variance, inherently determining the cut-off levels of the features, and not requiring data scaling and standardization in advance. We suggest the use of ALLvsNF model to discriminate between NF and CS in the screening testing stage (Stage-1) for the diagnosis of CS in clinical evaluations. Furthermore, we suggest the use of multiclass ALL model to discriminate among subtypes of CS in the follow-up testing stage (Stage-2) for the identification of the cause of CS in clinical evaluations. Also, multiclass ALL model with Stage-1 features can be employed to get an early opinion about the subtype of CS for diagnosis, prognosis, and treatment choices. The developed ML models outperformed the physicians' judgments under the constraint of using only the selected biochemical test findings utilized in ML model development. These suggest that ML approach can help improve physicians' judgment in diagnosing CS subtypes with limited biochemical tests which are cumbersome and stressful for patients.

Conflicts of Interest

The authors have no competing interests to disclose.

Acknowledgment

We thank Umit Cavdar, Dilek Karakus, Meric Coskun, Banu Erturk, and Mehmet Eren Kalender for their expertise in clinical evaluations.

References

[1] H. Cushing, *The pituitary body and its disorders: Clinical states produced by disorders of the hypophysis cerebri*. Philadelphia & London: J.B. Lippincott Company, 1912.

[2] L. K. Nieman, "Diagnosis of Cushing's Syndrome in the modern era," *Endocrinol. Metab. Clin. North Amer.*, vol. 47, no. 2, pp. 259–273, 2018.

[3] A. Lacroix, R. A. Feelders, C. A. Stratakis, and L. K. Nieman, "Cushing's Syndrome," *Lancet*, vol. 386, no. 9996, pp. 913–927, 2015.

[4] L. Vilar *et al.*, "Pitfalls in the diagnosis of Cushing's Syndrome," *Arquivos Brasileiros De Endocrinologia E Metabologia*, vol. 51, pp. 1207–1216, 2007.

[5] R. Alwani, L. Schmit Jongbloed, F. De Jong, A.-J. van der Lely, W. De Herder, and R. Feelders, "Differentiating between Cushing's disease and Pseudo-Cushing's Syndrome: Comparison of four tests," *Eur. J. Endocrinol.*, vol. 170, no. 4, pp. 477–86, 2014.

[6] R. Görges, G. Knappe, H. Gerl, M. Ventz, and F. Stahl, "Diagnosis of Cushing's Syndrome: Re-evaluation of midnight plasma cortisol vs urinary free cortisol and low-dose dexamethasone suppression test in a large patient group," *J. Endocrinol. Investigat.*, vol. 22, no. 4, pp. 241–249, 1999.

[7] J. Newell-Price and A. B. Grossman, "Differential diagnosis of Cushing's Syndrome," *Arquivos Brasileiros De Endocrinologia E Metabologia*, vol. 51, pp. 1199–1206, 11 2007.

[8] F. Holleman, E. Endert, M. Prummel, M. van Vessem-Timmermans, W. Wiersinga, and E. Fliers, "Evaluation of endocrine tests. B: Screening for hypercortisolism," *Neth. J. Med.*, vol. 63, no. 9, pp. 348–353, 2005.

[9] B. O. Åsvold, V. Grill, K. Thorstensen, and M. R. Bjørgaas, "Association between posttest dexamethasone and cortisol concentrations in the 1 mg overnight dexamethasone suppression test," *Endocr. Connections*, vol. 1, no. 2, pp. 62–67, 2012.

[10] F. Pecori Giraldi, A. G. Ambrogio, M. De Martin, L. M. Fatti, M. Scacchi, and F. Cavagnini, "Specificity of first-line tests for the diagnosis of Cushing's Syndrome: Assessment in a large series," *J. Clin. Endocrinol. Metab.*, vol. 92, no. 11, pp. 4123–4129, 2007.

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[12] E. R. Laws and M. P. Catalino, "Editorial. machine learning and artificial intelligence applied to the diagnosis and management of Cushing disease," *Neurosurg. Focus FOC*, vol. 48, no. 6, 2020, Art. no. E6.

[13] E. H. Wilkes, G. Rumsby, and G. M. Woodward, "Using machine learning to aid the interpretation of urine steroid profiles," *Clin. Chem.*, vol. 64, no. 11, pp. 1586–1595, 2018.

[14] J. Y. Yang *et al.*, "A hybrid machine learning-based method for classifying the Cushing's Syndrome with comorbid adrenocortical lesions," *BMC Genomic.*, vol. 9, no. S1, p. S 23, 2008.

[15] M. Zoli *et al.*, "Machine learning-based prediction of outcomes of the endoscopic endonasal approach in Cushing disease: Is the future coming?," *Neurosurg. Focus FOC*, vol. 48, no. 6, 2020, Art. no. E5.

[16] R. Wei *et al.*. M, "Deep-learning approach to automatic identification of facial anomalies in endocrine disorders," *Neuroendocrinol.*, vol. 110, no. 5, pp. 328–337, 2020.

[17] P. Schmitt, J. Mandel, and M. Guedj, "A comparison of six methods for missing data imputation," *J. Biometrics Biostatist.*, vol. 6, no. 1, p. 1, 2015.

[18] M. Hubert and S. Van der Veeken, "Robust classification for skewed data," *Adv. Data Anal. Classification*, vol. 4, no. 4, pp. 239–254, 2010.

[19] F. Siddiqui and Q. M. Ali, "Performance of non-parametric classifiers on highly skewed data," *Glob. J. Pure Appl. Math*, vol. 12, no. 2, pp. 1547–1565, 2016.

[20] D. P. Doane and L. E. Seward, "Measuring skewness: A forgotten statistic?," *J. Statist. Educ.*, vol. 19, no. 2, pp. 1–18, 2011.

[21] G. U. Yule and M. Kendall, "An introduction to the theory of statistics," 14th ed. London: C. Griffin Company, 1950.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[23] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.

[24] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain," *J. Clin. Epidemiol.*, vol. 54, no. 10, pp. 979–985, 2001.

[25] A. Tharwat, "Principal component analysis-a tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 3, pp. 197–240, 2016.

[26] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[28] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," *Proc. Int. Conf. Mach. Learn.*, vol. 96., 1996, pp. 148–156.

[29] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[30] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[31] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 431–439.

[32] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinf.*, vol. 7, no. 1, p. 91, 2006.

[33] B. van Calster, D. McLernon, M. van Smeden, L. Wynants, E. Steyerberg, and STRATOS Initiative, "Calibration: The achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, Dec. 2019.

[34] J. Hardt, M. Herke, T. Brian, and W. Laubach, "Multiple imputation of missing data: A simulation study on a binary response," *Open J. Statist.*, vol. 3, no. 5, pp. 370–378, 2013, doi: 10.4236/ojs.2013.35043.

[35] A. J. Masino *et al.*, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLoS One*, vol. 14, no. 2, 2019, Art. no. e0212665.

[36] M. R. Stavseth, T. Clausen, and J. Røoislien, "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data," *SAGE Open Med.*, vol. 7, 2019, Art. no. 2050312118822912.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[38] M. De Leo, A. Cozzolino, A. Colao, and R. Pivonello, "Subclinical Cushing's Syndrome," *Best Pract. Res. Clin. Endocrinol. Metab.*, vol. 26, no. 4, pp. 497–505, 2012.