Original Paper

Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study

Liyan Pan[1*], MSc; Guangjian Liu[1*], PhD; Xiaojian Mao[2*], PhD; Huixian Li[1], MSc; Jiexin Zhang[1], BD; Huiying Liang[1*], PhD; Xiuzhen Li[2*], PhD

[1]Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[2]Department of Genetics and Endocrinology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[*]these authors contributed equally

Corresponding Author:
Xiuzhen Li, PhD
Department of Genetics and Endocrinology
Guangzhou Women and Children's Medical Center
Guangzhou Medical University
9 Jinsui Road
Tianhe District
Guangzhou, 510623
China
Phone: 86 02038857692
Fax: 86 02038857692
Email: 13725100840@163.com

Abstract

Background: Central precocious puberty (CPP) in girls seriously affects their physical and mental development in childhood. The method of diagnosis—gonadotropin-releasing hormone (GnRH)–stimulation test or GnRH analogue (GnRHa)–stimulation test—is expensive and makes patients uncomfortable due to the need for repeated blood sampling.

Objective: We aimed to combine multiple CPP–related features and construct machine learning models to predict response to the GnRHa-stimulation test.

Methods: In this retrospective study, we analyzed clinical and laboratory data of 1757 girls who underwent a GnRHa test in order to develop XGBoost and random forest classifiers for prediction of response to the GnRHa test. The local interpretable model-agnostic explanations (LIME) algorithm was used with the black-box classifiers to increase their interpretability. We measured sensitivity, specificity, and area under receiver operating characteristic (AUC) of the models.

Results: Both the XGBoost and random forest models achieved good performance in distinguishing between positive and negative responses, with the AUC ranging from 0.88 to 0.90, sensitivity ranging from 77.91% to 77.94%, and specificity ranging from 84.32% to 87.66%. Basal serum luteinizing hormone, follicle-stimulating hormone, and insulin-like growth factor-I levels were found to be the three most important factors. In the interpretable models of LIME, the abovementioned variables made high contributions to the prediction probability.

Conclusions: The prediction models we developed can help diagnose CPP and may be used as a prescreening tool before the GnRHa-stimulation test.

(JMIR Med Inform 2019;7(1):e11728) doi:10.2196/11728

KEYWORDS

central precocious puberty; GnRHa-stimulation test; machine learning; prediction model

Introduction

Precocious puberty is related to the development of secondary sexual characteristics in girls before the age of 8 years and in boys before the age of 9 years. In recent years, the age of puberty onset has shown a decreasing trend, and puberty is related to subsequent health outcomes such as breast cancer, diabetes, and behavioral disorders [1]. Central precocious puberty (CPP), also known as true precocious puberty, is caused by early activation of the hypothalamic-pituitary-gonadal axis with clinical pubertal symptoms. CPP can influence final adult height and result in psychological problems, which will cause inappropriate behaviors. It is important for girls with suspected CPP to be evaluated and diagnosed in a timely manner.

The gold standard in the confirmation of CPP is the positive response of gonadotropin to a gonadotropin-releasing hormone (GnRH)–stimulation test. In the absence of GnRH, GnRH analogues (GnRHa) are usually used instead [2]. However, the stimulation test is time consuming and expensive. Besides, the test is painful and make patients uncomfortable due to the need for repeated blood sampling at different time points. Therefore, another method to avoid the disadvantages of the GnRHa-stimulation test will be of great help in the diagnosis of CPP.

Several studies have focused on investigating a single gonadotropin biomarker to identify patients with CPP conveniently. Basal or peak serum luteinizing hormone (LH), follicle-stimulation hormone (FSH), and the ratio LH/FSH are the most common biomarkers reported [3-7]. However, the cut-off values of these single biomarkers depend on the assay used to measure the gonadotropin levels. As a result, cut-off points in previous studies differed widely. Moreover, both the Pasternak group [3] and the Mogensen group [8] reported that a single basal serum LH measurement could verify the presence of CPP, but could not confirm the absence of CPP. Therefore, a single gonadotropin parameter alone may not be sufficient for the diagnosis of CPP, and clinical and laboratory factors that can predict response to the GnRHa-stimulation test should be combined [9,10]. Suh et al [10] found that accelerated growth rate, advanced bone age, and increased basal gonadotropin and insulin-like growth factor-I (IGF-I) levels were correlated with CPP. Traditional statistical analysis including *t* test and binary logistic regression were used to select factors correlated with the GnRH test [11-14]. Although remarkable progress has been made in these studies, there is a long way to go for their application in clinics due to the low sensitivity or specificity of tests.

Considering the previous studies and the extensive application of machine learning algorithms in the medical field, we aimed to determine whether combining multiple variables with machine learning classifiers could improve prediction of the GnRHa-stimulation test and thus help diagnose CPP.

Methods

Population and Variables

We enrolled 1757 girls with CPP symptoms who visited the Pediatric Day Ward of the Endocrinology Department at Guangzhou Women and Children's Medical Center from January 2012 to March 2018. All subjects had undergone the GnRHa-stimulation test. Girls with any other disorders or intracranial lesions were excluded from the study.

Girls fulfilling the following eligibility criteria were considered to have a positive response to the GnRHa-stimulation test and were diagnosed with CPP in our study: (1) peak LH level ≥ 10 IU/L or peak LH level ≥ 5 IU/L combined with a ratio of peak LH to FSH value ≥ 0.6 and (2) onset of secondary sexual characteristics at the age < 8 years. Girls whose laboratory tests did not satisfy all the abovementioned criteria were considered to have a negative response. According to the long-term clinical practices, the first condition with a peak LH ≥ 10 IU/L is used as the diagnosing criterion in our hospital. Peak LH level ≥ 5 IU/L combined with a ratio of peak LH to FSH value ≥ 0.6 is widely used in China and some other countries for children undergoing the GnRH-stimulation test [15,16]. Since the stimulation effect of GnRHa is almost hundreds times that of GnRH [17], a condition that affects the levels of sex hormones due to GnRH would do the same with GnRHa. Our diagnostic criteria are an improved version of the existing criteria that are adapted to our population.

Information such as chief complaints, development of secondary sexual characters, and abnormal duration of puberty were stored as free text in the clinical records of the electronic medical records system. Laboratory test values were reported as structured data in the laboratory information system. In total, 19 variables were extracted from the clinical records and laboratory results for all the 1757 patients. Specifically, 10 variables extracted from the clinical records were weight, height, body mass index (BMI), abnormal duration of puberty in records (History), menarche, core in breast (Core), pigmentation, development stage of pubes (Pubes), development stage of left breast, and development stage of right breast. Breast and pubic hair development were evaluated using Tanner staging (stages 1 to 5). Nine variables extracted from the laboratory results were age, basal serum LH, FSH, estradiol, prolactin, testosterone, growth hormone, IGF-I, IGF-binding protein-3 levels before the GnRHa test.

Among the 1757 patients, 436 girls had examination reports available, including pelvic ultrasonography (for development of the uterus and ovaries) and radiography of the left hand (for bone age). Six variables extracted from the examination reports were development of uterus, existence of follicle, uterine volume, left and right ovary volumes, and bone age. Bone age was measured by the Greulich and Pyle method [18]. The variables from the clinical records and the examination reports were extracted first with traditional regex match using Python [19] and then examined manually by two endocrinologists.

This study was approved by the Institutional Review Board of Guangzhou Women and Children's Medical Center and

conducted in accordance with the ethical guidelines of the Declaration of Helsinki of the World Medical Association. The requirement to obtain informed consent was waived because of the retrospective nature of the study. Data used in this study were anonymous, and no identifiable personal data of the patients were available for the analysis.

Data Preprocessing

Variables with more than 20% missing data, such as growth rate of height and weight and heights of parents, were excluded from this study. Missing values for continuous variables were replaced with mean values of all the samples grouped by age. Discrete variables like experience of menarche were filled with a value of 0. Discrete variables like Tanner stage for breast and pubes were filled with the least degree (stage 1).

Model Development and Assessment

Tree learning classifiers allow nonlinear interactions between features and have good interpretability. Considering this, we selected two tree-based ensemble binary classification algorithms—extreme gradient boosting (XGBoost) and random forest—to develop our models. We also used linear supported vector machines (SVM) and decision trees for the classification to compare the performance between ensemble models and nonensemble models. The models aimed to identify relationships between the input features and the output GnRHa test results, thereby determining whether a patient responds positively to the GnRHa test.

XGBoost is a scalable tree boosting and effective learning algorithm [20]. It trains a sequence of models to minimize errors made by existing models. Models in XGBoost are decision trees. Many data scientists have applied this algorithm to solve classification problems and achieved excellent results. XGBoost has also been successfully used in medical studies [21,22]. As XGBoost is essentially a gradient boosting tree model, which is not based on distance, normalization is not required. Random forest is another classical ensemble learning algorithm with a combination of a large amount of trees [23], which trains decision trees in parallel by using data with replacement. It applies bootstrap aggregating to tree learners, which leads to better performance as variance decreases. Random forest has the ability to handle nonlinear data and is robust to noise. Besides, parameter tuning is not that complex for these two algorithms compared to other ensemble learning algorithms. SVM is a binary classifier with a maximum margin hyperplane [24]. The decision tree classifier is a tree-like model used for classification [25].

In order to obtain robust assessments and prevent overfitting, we used a nested cross-validation with an outer Monte Carlo cross-validation [26] (MCCV, repeated 20 times) and an inner k-fold cross-validation (k=5) for parameter tuning, yielding a total of 20 times the five-fold cross-validation. In the outer MCCV loop, the whole data set is randomly divided into the training set (80%) and the test set (20%) for 20 times. For each training set, the inner stratified five-fold cross-validation loop is performed as follows. The training set is split into five subsets, where four subsets are used for training and one is used for test. Parameter tuning is performed with grid search in the inner cross-validation. Finally, a model fitted on the training set with parameters that has the best area under the curve (AUC) evaluated on the inner test set is determined. The detailed training and test process is presented in Figure 1.

Feature Importance

XGBoost and random forest classifier have the ability to evaluate the importance of features. Feature importance is a feature weight and can represent the contribution to prediction. In XGBoost, feature importance is computed by the sum of times that the feature is selected as a tree node. In random forest, feature importance is calculated based on the out-of-bag (OOB) error. OOB error is the mean prediction error for training observations in the respective bootstrap sample. After randomly adding noise perturbations to OOB samples, a feature with a higher OOB error difference is more important, with higher feature importance. For both models, each feature obtained 20 feature importance values with 20 times the MCCV. We summed all the feature importance values for each feature and obtained a rank for all features.

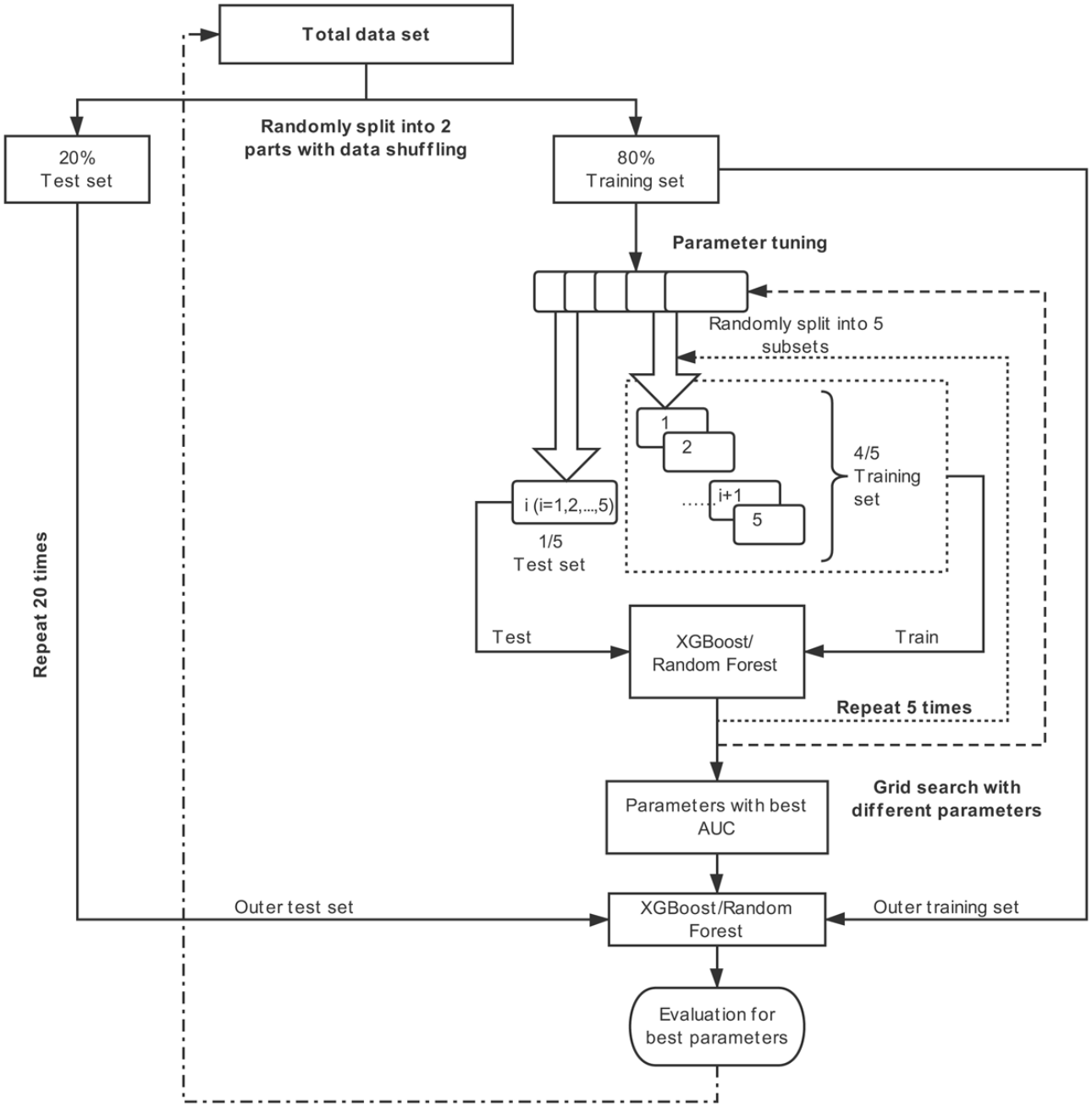
Model Interpretation

One disadvantage of machine learning is that the model usually runs as a black box. However, it is necessary for a doctor to understand the reasons why a model makes such a prediction in the clinic, especially when timely detection is necessary. Tree-based models can provide feature importance at a global level but not in a specific case. The local interpretable model-agnostic explanations (LIME) algorithm is developed to identify an interpretable model that is locally faithful for each individual prediction [27,28]. It provides relative feature contributions for a single instance of the prediction result. LIME generates neighborhood data by randomly perturbing features from the instance. It then learns locally weighted linear models on this neighborhood data to explain each of the classes in an interpretable way. Parameters in LIME mainly include the maximum number of features in explanation, number of neighborhood samples to generate, and machine learning prediction function. We used the LIME library from the original authors for the model interpretation. The number of neighborhood samples is 5000 by default. The parameter num_features (maximum number of features) is the number of features shown in the explanation. The default value is 10. In our study, the value was 9 for a clear layout, as the contributions of features ranked after 9 were almost zero.

In our study, we used a submodular pick [27] instead of a random pick to select a diverse, representative set of samples from the test set for nonredundant explanations. For these samples, we then obtained the class probabilities, and the representative individuals were assigned with and average weight contributed by each feature to display how the classifier made a decision. Finally, we went over all these LIME results together with the endocrinologists to decide whether we should trust the results of the model.

All computation and visualization were performed in Python [19] using packages like Scikit-learn, Pandas, Lime, and Matplotlib.

Figure 1. Training and validation process of prediction models. AUC: area under receiver operating characteristic; XGBoost: extreme gradient boosting.



Results

Subject Characteristics

Among the 1757 girls included in our study, 966 were positive for the GnRHa-stimulation test and diagnosed with CPP; the remaining 791 girls showed a negative response to the test. As shown in Table 1, 16 of the 19 variables were significantly higher in the CPP group than in the non-CPP group (P<.05), whereas prolactin, BMI, and pigmentation were similar in both groups.

Evaluation for Models

First, we developed prediction models with the data of 19 clinical and laboratory variables (Table 1) from all 1757 patients. Two machine learning algorithms, XGBoost and random forest classifiers, were used, and parameters with the best AUC were selected for each model. The performance as well as the selected parameters of the models are listed in Table 2, and their receiver operating characteristic (ROC) curves are plotted in Figure 2. The performance was evaluated on the 20 test sets split with MCCV. Both models had strong prediction powers, with a specificity of ≥84.32%, a sensitivity of ≥77.91%, and an AUC of ≥0.88. The XGBoost classifier is slightly more effective than the random forest classifier, especially in terms of the specificity (P<.01), whereas random forest is much more efficient in terms of the computation speed with less model complexity.

Table 1. Basic characteristics of girls who underwent the GnRHa-stimulation test.

Variables	Non-CPP[a] (n=966), mean (SD)	CPP (n=791), mean (SD)	*P* value[b]
Age (years)	7.07 (1.11)	7.52 (0.99)	<.001
LH[c] (IU/L)	0.12 (0.23)	0.93 (1.28)	<.001
FSH[d] (IU/L)	1.82 (1.30)	3.01 (1.62)	<.001
GH[e] (ng/mL)	3.27 (3.26)	4.75 (4.69)	<.001
IGF-I[f] (ng/mL)	231.35 (65.93)	317.87 (89.84)	<.001
IGFBP-3[g] (μg/mL)	4.55 (0.52)	4.81 (0.55)	<.001
Estradiol (pmol/L)	102.56 (50.96)	125.81 (60.97)	<.001
Prolactin (ng/mL)	8.73 (5.39)	8.59 (5.61)	.52
Testosterone (nmol/L)	0.80 (0.39)	0.94 (0.49)	<.001
History[h] (months)	7.67 (10.39)	9.27 (9.63)	<.001
Menstruation/menarche (yes, no)	N/A[i]	N/A	.03
Height[j] (cm)	127.16 (8.61)	131.61 (8.42)	<.001
Weight[j] (kg)	27.32 (5.32)	29.60 (4.95)	<.001
BMI[k] (kg/m^2)	16.73 (2.30)	16.91 (1.96)	.34
Breast core (yes, no)	N/A	N/A	.02
Pubes[l] (1-5)	1.06 (0.27)	1.14 (0.44)	<.001
Pigmentation (yes, no)	N/A	N/A	.87
Left breast[l] (1-5)	2.33 (0.84)	2.76 (0.92)	<.001
Right breast[l] (1-5)	2.32 (0.84)	2.78 (0.92)	<.001

[a]CPP: central precocious puberty.

[b]The equality of each indicator was evaluated by Chi-square or Student *t* test. *P*<.05 was considered significant.

[c]LH: luteinizing hormone.

[d]FSH: follicle-stimulation hormone.

[e]GH: growth hormone.

[f]IGF-I: insulin-like growth factor-I.

[g]IGFBP-3: insulin-like growth factor binding protein-3.

[h]Abnormal duration in records.

[i]N/A: not applicable.

[j]At stimulation test.

[k]BMI: body mass index.

[l]Tanner stage.

Table 2. Predictive performance of classifiers and the corresponding parameters. A paired *t* test was performed on specificity and sensitivity for comparison against XGBoost.

Algorithms/Variables	Specificity[a] (%), mean (SD)	Sensitivity[b] (%), mean (SD)	AUC[c], mean (SD)	Parameters
19 variables, 1757 patients				
XGBoost[d]	85.39 (1.38)	77.94 (3.50)	0.89 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
Random forest	84.32 (1.88)[e]	77.91 (3.59)[f]	0.88 (0.02)	Max depth=3, criterion=gini, number of trees=20
SVM[g]	88.94 (1.76)[e]	62.36 (4.12)[e]	0.86 (0.04)	Kernel=linear, penalty coefficient=5
Decision tree	75.90 (2.47)[e]	71.71 (3.99)[e]	0.74 (0.02)	Criterion=entropy
19 variables, 436 patients				
XGBoost	83.17 (5.29)	75.28 (6.43)	0.86 (0.04)	Learning rate=0.01, max depth=3, number of trees=500
Random forest	83.46 (6.28)[f]	74.72 (6.43)[f]	0.85 (0.04)	Max depth=3, criterion=gini, number of trees=20
SVM	88.94 (4.90)[e]	62.36 (7.73)[e]	0.86 (0.02)	Kernel=linear, penalty coefficient=5
Decision tree	76.25 (7.07)[e]	68.06 (7.12)[e]	0.72 (0.04)	Criterion=entropy
25 variables, 436 patients				
XGBoost[d]	87.66 (5.52)	76.64 (6.51)	0.90 (0.04)	Learning rate=0.01, max depth=4, number of trees=500
Random forest	87.41 (4.22)[f]	75.03 (7.91)[f]	0.90 (0.05)	Max depth=3, criterion=entropy, number of trees=20
SVM	89.81 (4.28)[f]	66.53 (7.01)[e]	0.86 (0.02)	Kernel=linear, penalty coefficient=5
Decision tree	76.35 (5.51)[e]	68.61 (7.16)[e]	0.72 (0.05)	Criterion=entropy
1-3 variables, 1757 patients, XGBoost[d]				
LH[h], IGF-I[i], FSH[j]	83.17 (1.62)	76.39 (3.57)	0.86 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
LH[h], IGF-I[i]	83.27 (1.62)	75.69 (3.61)	0.86 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
LH[h], FSH[j]	83.56 (1.94)	75.83 (3.13)	0.84 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
LH[h]	83.37 (2.00)	75.97 (3.74)	0.84 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
IGF-I[i], FSH[j]	80.77 (2.47)	57.08 (3.29)	0.77 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
IGF-I[i]	80.19 (3.14)	53.19 (4.55)	0.73 (0.02)	Learning rate=0.01, max depth=3, number of trees=500
FSH[j]	84.13 (3.87)	45.00 (5.34)	0.68 (0.02)	Learning rate=0.01, max depth=3, number of trees=500

[a]Specificity=number of true negatives/(number of true negatives+number of false positives).

[b]Sensitivity=number of true positives/(number of true positives+number of false negatives).

[c]AUC, area under the receiver operating curve.

[d]XGBoost: extreme gradient boosting.

[e]$P<.01$

[f]Not significant.

[g]SVM: supported vector machines.

[h]LH: luteinizing hormone.

[i]IGF-I: insulin-like growth factor-I.

[j]FSH: follicle-stimulation hormone.

In the data set, 436 girls had additional examination reports, and we extracted six variables from these reports (see Population and Variables subsection). To investigate whether adding image features could enhance the prediction efficiency, we combined the six variables with the 19 variables and trained and evaluated both the XGBoost and random forest models on the 436 samples, of which 180 patients belonged to the CPP group and 256 belonged to the non-CPP group. For the ease of comparison, we retrained the previous 19-variable models with the 436 samples. As shown in Table 2, the reduction in sample size led to a serious decline in model performance, whereas the addition of six image features improved their performance. Specifically, for XGBoost in 436 samples, the specificity increased from 83.17% for 19 variables to 87.66% for 25 variables, the

sensitivity increased from 75.28% to 76.64%, and the AUC increased from 0.86 to 0.90. For random forest in 436 samples with 25 variables, the specificity increased from 83.46% to 87.41%, the sensitivity increased from 74.72% to 75.03%, and the AUC increased from 0.86 to 0.90 compared to the results from 436 samples with 19 variables. Similarly, as seen in the ROC curves shown in Figure 2, XGBoost performed slightly better than the random forest classifier.

To compare performance between ensemble models and nonensemble models, the SVM and decision tree classifiers were used to develop predictive models for the abovementioned settings. Higher specificities were achieved with the SVM models. However, sensitivities for the SVM models were much lower than those for the ensembles models. The decision tree models demonstrated significantly inferior performance in terms of almost all the sensitivities, specificities, and AUCs. These results suggest that the ensemble models are able to yield excellent performance while maintaining a good balance between sensitivity and specificity in the prediction of CPP.

Feature Importance

We computed the feature importance score for all the 19 variables to identify important features used by the models. The importance of each feature calculated by the models is plotted in Figure 3. In both models, the most important predictive variable was LH level, followed by IGF-I and FSH levels. The fourth most important feature for the random forest model was height, which ranked fifth in the XGBoost model. Prolactin is the fourth most important feature of XGBoost, but it contributed only a little to random forest. These data suggest that different machine learning algorithms attach importance to different combination lists of variables, although they yield similar predictive performance.

In order to further verify the importance of the top 3 features, we constructed XGBoost models with these features individually or in combination (Table 2). As expected, the models using one, two, or three features had poorer performances than the models using all features. The results showed that the higher the feature ranked, the better the corresponding model performed. Interestingly, LH alone or together with IGF-I and FSH is sufficient to predict a response to the GnRHa test with a fairly good performance and an AUC between 0.84 and 0.86. These data support the results from the feature importance calculations.

Figure 2. ROC curves for classifiers with 19 variables for 1757 patients and 25 variables for 436 patients. ROC: receiver operating curve; AUC: area under ROC.

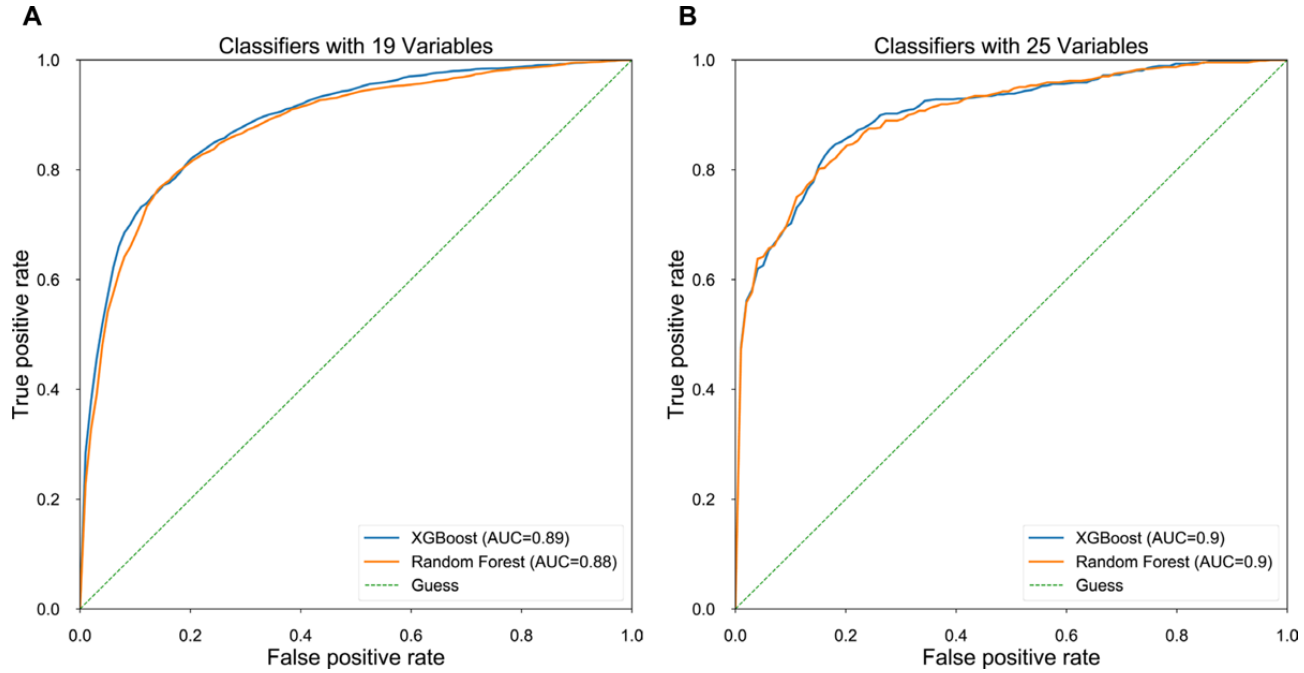
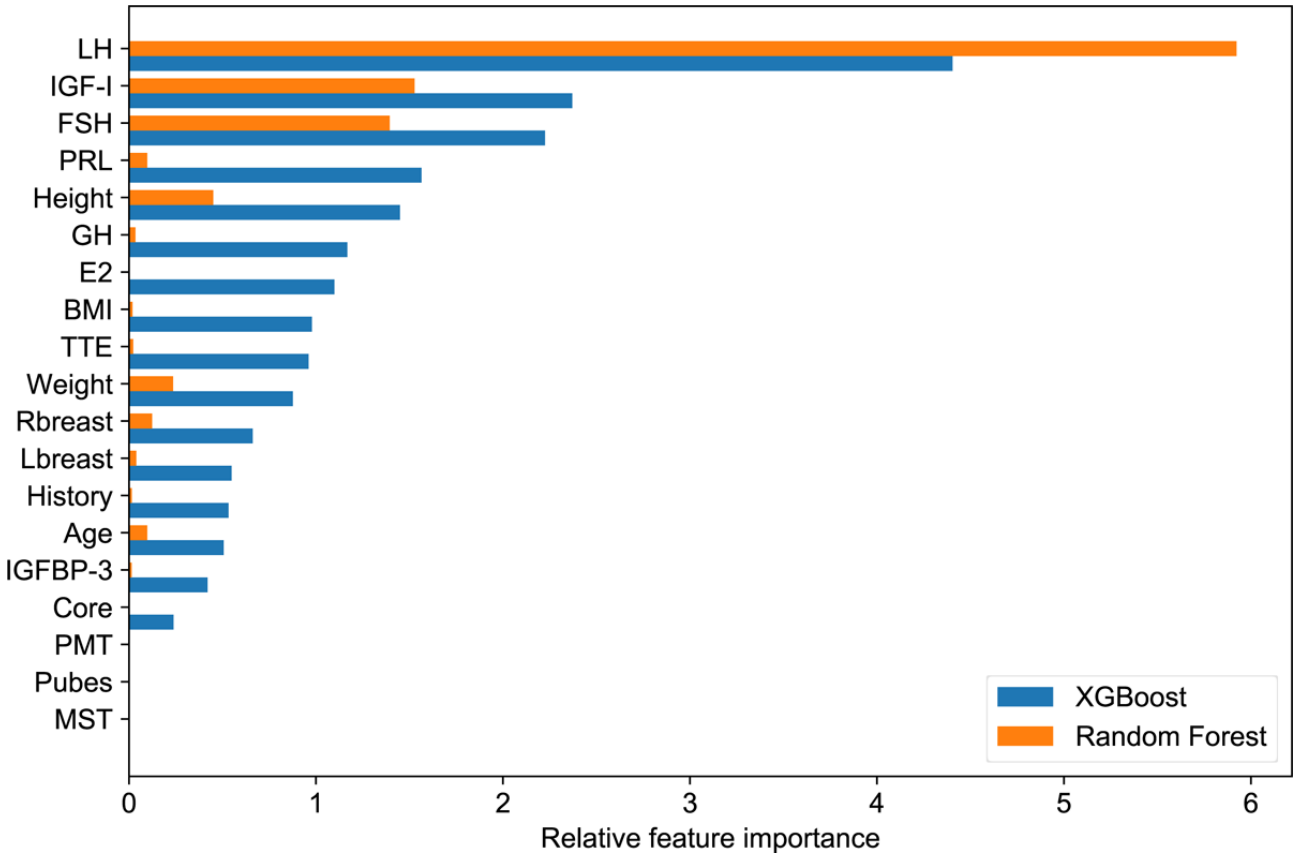


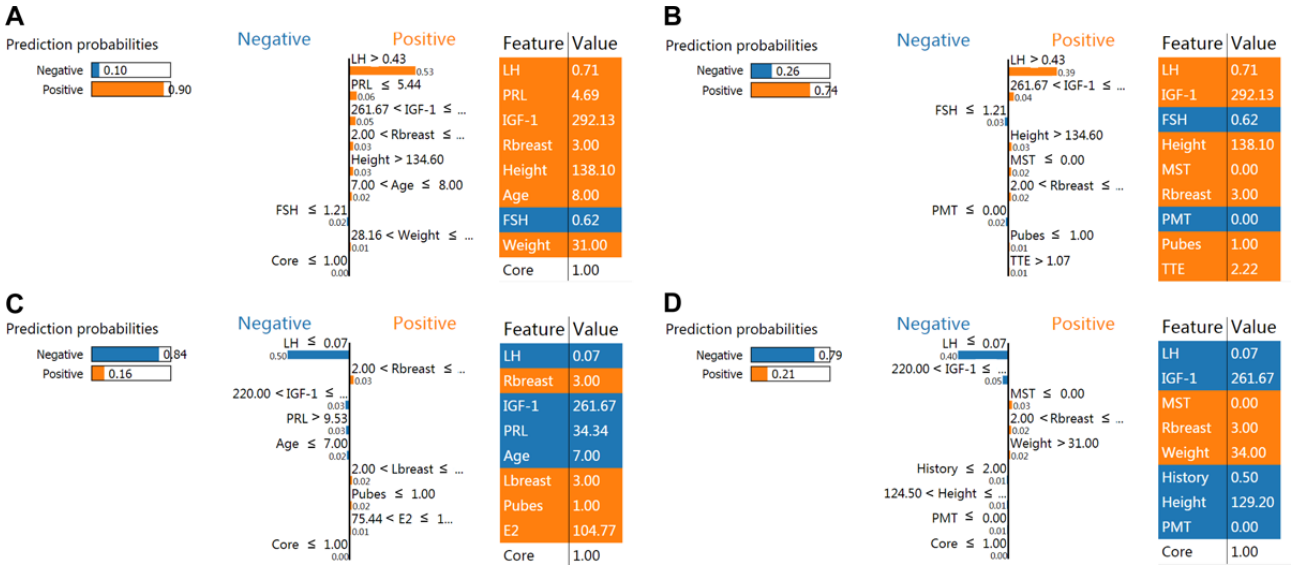
Figure 3. Feature importance ranking for 19 variables in two classifiers calculated by the models. LH: luteinizing hormone; IGF-I: insulin-like growth factor-I; FSH: follicle-stimulation hormone; PRL: prolactin; GH: growth hormone; E2: estradiol; BMI: body mass index; TTE: testosterone; Rbreast: right breast; Lbreast: left breast; IGFBP-3: insulin-like growth factor binding protein-3; PMT: pigmentation; MST: menstruation.



Local Interpretable Model-Agnostic Explanations for Interpretation

A representative set of 200 samples, which accounted for more than 50% of the test set and were enough to be the budget size of individual instances to understand a model, were selected with the submodular pick method [27] for the 19-variable models. LIME was then applied to investigate feature contributions for each prediction. Results with top 9 features are presented in Figure 4 for one positive sample and one negative sample (more representative samples can be seen in Multimedia Appendix 1). In Figure 4, XGBoost predicts an instance where CPP positively responds to the GnRHa test with a probability of 90%. Only the feature growth hormone supports the negative prediction, whereas LH, prolactin, IGF-I, and information about body development support the positive prediction. This makes sense in the clinical diagnosis of CPP and reveals that we can trust our prediction models to a certain extent. In Figure 4, left and right breast at Tanner stage 3, FSH level > 2 IU/L, and several other features support the positive prediction with a probability of 16%. LH level of 0.07 IU/L, prolactin level > 9.53 ng/mL, IGF-I level > 220 ng/mL, and age < 7 years contribute to the negative prediction with a probability of 84%. Similar results are observed in Figure 4.

Figure 4. Results of LIME with XGBoost and Random Forest classifiers applied to one positive (A, B) and one negative (C, D) instance. The left sides are for XGBoost, and the right for Random Forest. Blue color is for the negative instance and orange is for the positive instance. The first column represents the prediction probabilities of negative and positive results achieved from classifiers. The second column shows the features' contributions to the probability. Only the top nine features are displayed for clarity. The third column displays the original data values. LIME: local interpretable model-agnostic explanations; XGBoost: extreme gradient boosting; LH: luteinizing hormone; IGF-I: insulin-like growth factor-I; FSH: follicle-stimulation hormone; PRL: prolactin; GH: growth hormone; E2: estradiol; BMI: body mass index; TTE: testosterone; Rbreast: right breast; Lbreast: left breast; IGFBP-3: insulin-like growth factor binding protein-3; PMT: pigmentation; MST: menstruation; PMT: pigmentation.



Discussion

Overview

CPP mimics pubertal development ahead of time at an inappropriate chronological age. It requires timely detection and treatment in case of physical and physiological effect on girls. The GnRHa-stimulation test is expensive and time consuming and causes discomfort to patients. Here, we applied machine learning algorithms to multiple clinical variables and built two tree-based ensemble learning classifiers for the prediction of response to the GnRHa-stimulation test. Both the XGBoost and random forest models achieved good performance in distinguishing between positive and negative responses, with the AUC ranging from 0.88 to 0.90, the sensitivity ranging from 77.91% to 77.94%, and the specificity ranging from 84.32% to 87.66%.

Comparisons with Previous Models

Several previous models focused on determining optimal blood sampling time points or appropriate cut-off values to simplify the stimulation test. Kandemir et al [29] found that a single sample of LH tested at the 40th minute after stimulation with a cut-off of 5 IU/L could yield 98% sensitivity and 100% specificity in the diagnosis of CPP. Yazdani et al [30] showed that an LH concentration > 5 IU/L at 3 hours has optimal sensitivity (83%) and specificity (97%). In the study of Çatlı et al [7], 100% sensitivity and 84% specificity were obtained using a cut-off value > 0.24 for peak LH/FSH ratio in girls. Although these models performed better than our models, they had to be used after stimulation and therefore could not avoid the disadvantages of the GnRH/GnRHa test completely.

Some models used only the basal sex hormone level. Yazdani et al [30] found that a basal LH level of >0.1 IU/L, a basal LH/FSH ratio >1, and basal estradiol level ≥1.5 ng/dL in girls have low sensitivity (10%-67%) but excellent specificity (94%-100%). Çatlı et al [7] also reported models with the basal FSH or LH levels and achieved a sensitivity of 71% and a specificity of 68% or 64%. Pasternak et al [3] reported that basal LH levels ≤ 0.1 IU/L were sufficient to rule out positive response to the GnRH test with a specificity of 94% but a sensitivity of only 64% in girls. In another model [4], the basal LH level with a cut-off value of 0.35 IU/L was associated with a sensitivity of 63.96% and a specificity of 76.3% based on the ROC with an AUC of 0.77. These results varied a lot due to the different settings and sample sizes. In this study, our models showed better performance with more features before stimulation and a larger homogeneous population, which is the largest population in such a study to our knowledge.

Predictive Features

Based on our machine learning models, basal LH, IGF-I, and FSH levels are predictive factors with top ranks for the feature importance in both models. Previous studies have demonstrated that the measurement of LH could be better than that of other sex hormones for initial evaluation of suspected puberty [3,8]. In our study, the LH level ranked first and was much more important than other variables. Besides LH, another indicator monitored in the stimulation test, FSH, was also selected by the models as the third most–important variable. Obviously, LH and FSH are important to CPP because they are biomarkers of the hypothalamic-pituitary-gonadal axis activation, which is the essence of CPP. IGF-I, which is the second most important variable in our models, is reportedly involved in GnRH regulation [31,32] and is increasingly expressed in girls with CPP [9,5]. Animal studies showed that the IGF-I signaling pathways play important roles in the timing of puberty in girls [32]. Although IGF-I has not been considered in previous models, our study suggests that IGF-I may be a valuable marker for diagnosing CPP.

Several studies [12,33,34] suggested that image reports like pelvic ultrasound and radiography of the hand have adjunct diagnostic values in CPP diagnosis but provide no reliable differentiation alone. Here, we found that adding features from the image reports improved the prediction results. Performance of models built based on 1757 samples was better than that based on 436 samples, suggesting a sample size effect. Interestingly, in the case of 436 samples with additional six image variables, the aforementioned sample size effect was balanced. This suggests that more samples with image features will produce better results. Thus, medical image examinations like bone age radiography should be considered before the GnRHa-stimulation test for girls with suspected CPP.

Interpretations of Models

We noticed that more variables were assigned with a moderate value of feature importance in the XGBoost model than those in the random forest model. This is reasonable when considering the different algorithms the two models used for prediction and importance evaluation. In XGBoost, trees are sequentially built in a boosting manner to enhance the overall performance. The estimates of feature importance are provided explicitly with the frequency that the feature is selected as a tree node from a trained predictive model. In contrast, trees are trained parallelly in a bootstrapping way in random forest to vote for the final decision. The feature importance is estimated implicitly through permuting the feature's values and calculating the change of the model's prediction error. Obviously, XGBoost includes each contribution of each feature to each tree into the feature importance, whereas random forest only evaluates each feature globally without specific contributions. It should be noted that different combinations of variables may produce models with similar predictive accuracy, relating to the uncertainty analysis of the solutions in any decision-making problem [35-37]. This is not rare in machine learning models in medicine [38-40]. Moreover, the most important features in the clinic such as basal LH, IGF-I, and FSH levels were all sorted out by both models, demonstrating that they are both reliable and effective in predicting response to the GnRHa test.

In order to provide endocrinology physicians a trustworthy insight into the prediction models, we also used LIME to show each feature's contribution to predicting probabilities reasonably. The most important features used by the models for individual prediction have been proven to be significant in the clinic [3-5,7,9,10], demonstrating that our models are credible. This will greatly increase the interpretability of the machine learning models and make it convenient for individualized diagnosis in the clinic.

Limitations

There are some limitations to this study. First, growth velocity is specially related to physical development. Due to the lack of height growth rate and weight growth rate, we did not include growth velocity in our feature set. For further research, we will focus more on medical imaging and growth velocity to identify their diagnostic value with CPP. Second, our work included only girls with suspected CPP from a single center in China. The prediction models in this study may not be suitable for the population in other districts or countries. Third, manual inspection of values extracted through regular expression matching from free text could reduce errors to improve the model performance. However, this adds a considerable amount of work and thus reduces the scalability of the model. We are improving the matching algorithm with the manually inspected data to increase the level of model automation. Finally, features generated from laboratory results are more complete than those extracted from free text, which may affect the rank of feature importance. More efforts are required to enhance the data quality of unstructured features in the future.

Conclusions

Our study is the first one to apply both machine learning algorithms and the explanation method to the diagnosis of CPP. Our models can predict the response to the stimulation test before injection of GnRHa in girls who are suspected of having CPP and thus may be used as a prescreening tool to help physicians make decisions in conjunction with the GnRHa-stimulation test.

Acknowledgments

This work was supported by Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center (Grant #KCP-2016-002).

Authors' Contributions

HL and XL proposed the project and provided supervision. LP is responsible for data collection, project implementation, and manuscript drafting and revision. GL implemented the project and performed manuscript editing. XM provided clinical guidance and revised the manuscript. HL conducted part of the data collection and statistical analysis. JZ helped collect part of the data. All authors reviewed the manuscript in its final form.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Results of LIME with XGBoost and Random Forest classifiers applied to four positive and four negative instances. LIME: local interpretable model-agnostic explanations.

[PDF File (Adobe PDF File), 267KB - medinform_v7i1e11728_app1.pdf]

References

1. Abreu AP, Kaiser UB. Pubertal development and regulation. Lancet Diabetes Endocrinol 2016 Mar;4(3):254-264 [FREE Full text] [doi: 10.1016/S2213-8587(15)00418-0] [Medline: 26852256]
2. Houk C, Kunselman AR, Lee PA. The diagnostic value of a brief GnRH analogue stimulation test in girls with central precocious puberty: a single 30-minute post-stimulation LH sample is adequate. J Pediatr Endocrinol Metab 2008 Dec;21(12):1113-1118. [Medline: 19189683]
3. Pasternak Y, Friger M, Loewenthal N, Haim A, Hershkovitz E. The utility of basal serum LH in prediction of central precocious puberty in girls. Eur J Endocrinol 2012 Feb;166(2):295-299 [FREE Full text] [doi: 10.1530/EJE-11-0720] [Medline: 22084156]
4. Ding Y, Li J, Yu Y, Yang P, Li H, Shen Y, et al. Evaluation of basal sex hormone levels for activation of the hypothalamic-pituitary-gonadal axis. J Pediatr Endocrinol Metab 2018 Mar 28;31(3):323-329. [doi: 10.1515/jpem-2017-0124] [Medline: 29466239]
5. Sørensen K, Aksglaede L, Petersen JH, Andersson AM, Juul A. Serum IGF1 and insulin levels in girls with normal and precocious puberty. Eur J Endocrinol 2012 May;166(5):903-910. [doi: 10.1530/EJE-12-0106] [Medline: 22379117]
6. Houk C, Kunselman AR, Lee PA. Adequacy of a single unstimulated luteinizing hormone level to diagnose central precocious puberty in girls. Pediatrics 2009 Jun;123(6):e1059-e1063. [doi: 10.1542/peds.2008-1180] [Medline: 19482738]
7. Çatlı G, Erdem P, Anık A, Abacı A, Böber E. Clinical and laboratory findings in the differential diagnosis of central precocious puberty and premature thelarche. Turk Pediatri Ars 2015 Mar;50(1):20-26 [FREE Full text] [doi: 10.5152/tpa.2015.2281] [Medline: 26078693]
8. Mogensen SS, Aksglaede L, Mouritsen A, Sørensen K, Main KM, Gideon P, et al. Diagnostic work-up of 449 consecutive girls who were referred to be evaluated for precocious puberty. J Clin Endocrinol Metab 2011 May;96(5):1393-1401. [doi: 10.1210/jc.2010-2745] [Medline: 21346077]
9. Nam H, Rhie YJ, Son CS, Park SH, Lee K. Factors to predict positive results of gonadotropin releasing hormone stimulation test in girls with suspected precocious puberty. J Korean Med Sci 2012 Feb;27(2):194-199 [FREE Full text] [doi: 10.3346/jkms.2012.27.2.194] [Medline: 22323868]
10. Suh J, Choi MH, Kwon AR, Kim YJ, Jeong JW, Ahn JM, et al. Factors that predict a positive response on gonadotropin-releasing hormone stimulation test for diagnosing central precocious puberty in girls. Ann Pediatr Endocrinol Metab 2013 Dec;18(4):202-207 [FREE Full text] [doi: 10.6065/apem.2013.18.4.202] [Medline: 24904878]
11. Xu Y, Li GM, Li Y. Advanced bone age as an indicator facilitates the diagnosis of precocious puberty. J Pediatr (Rio J) 2018;94(1):69-75 [FREE Full text] [doi: 10.1016/j.jped.2017.03.010] [Medline: 28866322]
12. Lee SH, Joo EY, Lee J, Jun Y, Kim M. The Diagnostic Value of Pelvic Ultrasound in Girls with Central Precocious Puberty. Chonnam Med J 2016 Jan;52(1):70-74 [FREE Full text] [doi: 10.4068/cmj.2016.52.1.70] [Medline: 26866003]
13. Lee H, Choi S, Fan D, Jang K, Kim M, Hwang C. Evaluation of characteristics of the craniofacial complex and dental maturity in girls with central precocious puberty. Angle Orthod 2018 Apr 30. [doi: 10.2319/112317-809.1] [Medline: 29708396]
14. Baik JS, Choi JW, Kim SJ, Kim JH, Kim S, Kim JH. Predictive Value of Dental Maturity for a Positive Gonadotropin-Releasing Hormone Stimulation Test Result in Girls with Precocious Puberty. J Korean Med Sci 2017 Feb;32(2):296-302 [FREE Full text] [doi: 10.3346/jkms.2017.32.2.296] [Medline: 28049241]
15. Subspecialty Group of Endocrinologic, Hereditary and Metabolic Diseases, the Society of Pediatrics, Chinese Medical Association, Editorial Board, Chinese Journal of Pediatrics. [Consensus statement For the diagnosis and treatment of central precocious puberty (2015)]. Zhonghua Er Ke Za Zhi 2015 Jun;53(6):412-418. [Medline: 26310550]
16. Carel J, Eugster EA, Rogol A, Ghizzoni L, Palmert MR, ESPE-LWPES GnRH Analogs Consensus Conference Group, et al. Consensus statement on the use of gonadotropin-releasing hormone analogs in children. Pediatrics 2009 Apr;123(4):e752-e762. [doi: 10.1542/peds.2008-1783] [Medline: 19332438]
17. Ibáñez L, Potau N, Zampolli M, Virdis R, Gussinyé M, Carrascosa A, et al. Use of leuprolide acetate response patterns in the early diagnosis of pubertal disorders: comparison with the gonadotropin-releasing hormone test. J Clin Endocrinol Metab 1994 Jan;78(1):30-35. [doi: 10.1210/jcem.78.1.7507123] [Medline: 7507123]
18. Bayley N, Pinneau SR. Tables for predicting adult height from skeletal age: Revised for use with the greulich-pyle hand standards. The Journal of Pediatrics 1952 Apr;40(4):423-441. [doi: 10.1016/S0022-3476(52)80205-7]
19. Python. URL: https://www.python.org/ [accessed 2019-01-22] [WebCite Cache ID 75cTVrNhZ]
20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016 Aug 13 Presented at: KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA, USA. [doi: 10.1145/2939672.2939785]
21. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 2017 Sep;4(3):159-169 [FREE Full text] [doi: 10.1007/s40708-017-0065-7] [Medline: 28434153]

22. Yao D, Calhoun VD, Fu Z, Du Y, Sui J. An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. J Neurosci Methods 2018 May 15;302:75-81. [doi: 10.1016/j.jneumeth.2018.03.008] [Medline: 29578038]
23. Liaw A, Wiener M. R News. 2002 Dec. Classification and regression by randomForest URL: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf [accessed 2019-01-22] [WebCite Cache ID 75cYZwTh7]
24. Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998 Jul 10;13(4):18-28. [doi: 10.1109/5254.708428] [Medline: 21889629]
25. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660-674. [doi: 10.1109/21.97458]
26. Xu Q, Liang Y. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems 2001 Apr;56(1):1-11. [doi: 10.1016/S0169-7439(00)00122-2]
27. Ribeiro MT, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier. 2016 Aug Presented at: KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. E. [doi: 10.1145/2939672.2939778]
28. Liu N, Kumara S, Reich E. Explainable data-driven modeling of patient satisfaction survey data. : IEEE; 2017 Dec Presented at: 2017 IEEE International Conference on Big Data; 11-14 Dec. 2017; Boston, MA, USA. [doi: 10.1109/BigData.2017.8258391]
29. Kandemir N, Demirbilek H, Özön ZA, Gönç N, Alikaşifoğlu A. GnRH stimulation test in precocious puberty: single sample is adequate for diagnosis and dose adjustment. J Clin Res Pediatr Endocrinol 2011 Mar;3(1):12-17 [FREE Full text] [doi: 10.4274/jcrpe.v3i1.03] [Medline: 21448328]
30. Yazdani P, Lin Y, Raman V, Haymond M. A single sample GnRHa stimulation test in the diagnosis of precocious puberty. Int J Pediatr Endocrinol 2012 Jul 18;2012(1):23 [FREE Full text] [doi: 10.1186/1687-9856-2012-23] [Medline: 22809285]
31. Wolfe A, Divall S, Wu S. The regulation of reproductive neuroendocrine function by insulin and insulin-like growth factor-1 (IGF-1). Front Neuroendocrinol 2014 Oct;35(4):558-572 [FREE Full text] [doi: 10.1016/j.yfrne.2014.05.007] [Medline: 24929098]
32. Divall SA, Williams TR, Carver SE, Koch L, Brüning JC, Kahn CR, et al. Divergent roles of growth factors in the GnRH regulation of puberty in mice. J Clin Invest 2010 Aug;120(8):2900-2909 [FREE Full text] [doi: 10.1172/JCI41069] [Medline: 20628204]
33. de Vries L, Phillip M. Role of pelvic ultrasound in girls with precocious puberty. Horm Res Paediatr 2011 Feb;75(2):148-152. [doi: 10.1159/000323361] [Medline: 21228561]
34. Eksioglu A, Yilmaz S, Cetinkaya S, Cinar G, Yildiz YT, Aycan Z. Value of pelvic sonography in the diagnosis of various forms of precocious puberty in girls. J Clin Ultrasound 2013 Feb;41(2):84-93. [doi: 10.1002/jcu.22004] [Medline: 23124596]
35. Fernández Martínez J, Fernández Muñiz M, Tompkins M. On the topography of the cost functional in linear and nonlinear inverse problems. Geophysics 2012 Jan;77(1):W1-W15. [doi: 10.1190/geo2011-0341.1]
36. Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag 2006 Sep;6(3):21-45. [doi: 10.1109/MCAS.2006.1688199]
37. Fernández-Martínez J, Fernández-Muñiz Z, Pallero J, Pedruelo-González L. From Bayes to Tarantola: New insights to understand uncertainty in inverse problems. Journal of Applied Geophysics 2013 Nov;98:62-72. [doi: 10.1016/j.jappgeo.2013.07.005]
38. deAndrés-Galiana EJ, Fernández-Martínez JL, Luaces O, Del CJJ, Huergo-Zapico L, Acebes-Huerta A, et al. Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems. J Biomed Inform 2016 Apr;60:342-351 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.017] [Medline: 26956213]
39. Kosmicki J, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry 2015 Feb 24;5:e514 [FREE Full text] [doi: 10.1038/tp.2015.7] [Medline: 25710120]
40. Liu G, Xu Y, Wang X, Zhuang X, Liang H, Xi Y, et al. Developing a Machine Learning System for Identification of Severe Hand, Foot, and Mouth Disease from Electronic Medical Record Data. Sci Rep 2017 Nov 27;7(1):16341 [FREE Full text] [doi: 10.1038/s41598-017-16521-z] [Medline: 29180702]

Abbreviations

AUC: area under receiver operating characteristic
BMI: body mass index
CPP: central precocious puberty
FSH: follicle-stimulation hormone
GnRH: gonadotropin releasing hormone
GnRHa: gonadotropin releasing hormone analogues
IGF-I: insulin-like growth factor-I
IGFBP-3: insulin-like growth factor binding protein-3

LH: luteinizing hormone
LIME: local interpretable model-agnostic explanations
MCCV: Monte Carlo cross-validation
OOB: out-of-bag
ROC: receiver operating characteristic
SVM: supported vector machines
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 30.07.18; peer-reviewed by P Cai, J Chen; comments to author 06.09.18; revised version received 10.10.18; accepted 09.12.18; published 12.02.19

Please cite as:
Pan L, Liu G, Mao X, Li H, Zhang J, Liang H, Li X
Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study
JMIR Med Inform 2019;7(1):e11728
URL: http://medinform.jmir.org/2019/1/e11728/
doi:10.2196/11728
PMID:30747712