CLINICAL RESEARCH

e-ISSN 1643-3750
© Med Sci Monit, 2026; 32: e953235
DOI: 10.12659/MSM.953235

Received: 2026.02.26
Accepted: 2026.04.22
Available online: 2026.05.12
Published: 2026.XX.XX

Development and Validation of Machine-Learning-Based Prediction Models for Thyroid Diseases During Pregnancy

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

BCDEF Guang Yang
BCEF Yi Gao
BC Pengfei Liu
BC Jingwen Jiang
ADEG Hui Qiao
ACDE Weixuan Sheng

Department of Anesthesiology, Beijing Shijitan Hospital, Capital Medical University, Beijing, PR China

Corresponding Authors: Hui Qiao, Department of Anesthesiology, Beijing Shijitan Hospital, Capital Medical University, 10 Tieyi Road, Yangfangdian, Haidian District, Beijing, China, Phone: +86 15301063775, e-mail: qiao-hui240@163.com; Weixuan Sheng, Department of Anesthesiology, Beijing Shijitan Hospital, Capital Medical University, 10 Tieyi Road, Yangfangdian, Haidian District, Beijing, China, Phone: +86 15210644881, e-mail: swx0214@126.com

Financial support: None declared
Conflict of interest: None declared

Background: International guidelines recommend early screening based on targeted risk factors to identify thyroid disease during pregnancy. The complexity of these risk factors makes accurate prediction challenging. This study aimed to develop and compare multiple machine-learning-based predictive models for thyroid disease during pregnancy.

Material/Methods: This retrospective study analyzed the clinical characteristics of 5461 women who gave birth at a single center. The dataset was divided into training and test sets. In the training set, feature variables associated with thyroid disease during pregnancy were selected using the Boruta algorithm. Eight models were developed: logistic regression, Bayesian approach, k-nearest neighbors, support vector machine, neural network, classification and regression tree, extreme gradient boosting, and random forest (RF). Model performance was evaluated using the receiver operating characteristic (ROC) curve, precision-recall curve (PRC), calibration curve, and decision curve analysis.

Results: Nine feature variables were identified: age, height, pre-pregnancy weight, gravidity, parity, primiparity or multiparity, hypertensive disorders of pregnancy, scarred uterus, and autoimmune disease. The RF model demonstrated the best performance, with accuracy of 0.98387819 and 0.99597990, Matthews correlation coefficient of 0.96794139 and 0.97781292, log loss of 0.12670703 and 0.09442025, Brier score of 0.02495798 and 0.01921069, area under the ROC curve of 0.99877170 and 0.99991140, and area under the PRC of 0.99864486 and 0.99922572 in the training and test sets, respectively.

Conclusions: The RF model demonstrates excellent discriminative performance, accuracy, consistency, and generalizability in predicting thyroid disease during pregnancy.

Keywords: Machine Learning • Predictive Learning Models • Pregnancy Complications • Random Forest • Thyroid Diseases

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/953235

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Introduction

Thyroid disease during pregnancy is a serious condition affecting maternal and fetal health, primarily including hypothyroidism (both clinical and subclinical), hyperthyroidism, and pregnancy-specific physiological or pathological thyroid changes [1]. Subclinical hypothyroidism has an estimated prevalence of approximately 2% to 5% and is characterized by elevated thyrotropin (TSH) levels with normal free thyroxine (FT4) levels. Clinical hypothyroidism occurs in about 0.3% to 0.5% of pregnancies and is defined by elevated TSH levels (exceeding pregnancy-specific reference ranges) and decreased FT4 levels. Hyperthyroidism during pregnancy has a prevalence of approximately 0.1% to 0.4%, and Graves' disease represents around 85% of cases; the remainder are attributed to gestational transient thyrotoxicosis [2,3]. Risk factors vary across subtypes of thyroid disease during pregnancy. High-risk factors for hypothyroidism and subclinical hypothyroidism include positive thyroid peroxidase antibodies (TPOAb)—which substantially increase the risks of miscarriage and preterm birth—a personal or family history of thyroid disease, iodine deficiency or excess, obesity, and metabolic diseases such as type 1 diabetes. In contrast, high-risk factors for hyperthyroidism include a personal or family history of Graves' disease, multiple pregnancies associated with elevated human chorionic gonadotropin levels that may precipitate gestational thyroid dysfunction, and severe pregnancy-related vomiting, which is also linked to increased human chorionic gonadotropin levels [4-6]. There is evidence that early screening and appropriate intervention in high-risk groups can greatly improve maternal and fetal outcomes [7,8]. Although international guidelines recommend targeted screening based on high-risk factors to identify thyroid dysfunction during pregnancy, the complexity of these factors increases the difficulty of establishing an early prediction system [9].

The management of thyroid disease during pregnancy relies on early identification, risk stratification, and individualized interventions [2,10]. Conventional diagnostic methods such as TSH and FT4 testing, although effective, have limitations, including delayed detection (eg, subclinical hypothyroidism may be asymptomatic) and individual variability (eg, changes in pregnancy-specific reference ranges for TSH) [2,8]. Machine learning and predictive models show broad potential concerning thyroid disease during pregnancy by optimizing screening, diagnosis, treatment, and complication prediction, given their ability to integrate multidimensional data and achieve high accuracy in disease prediction, classification, and management [11,12]. Recent advances in machine learning have applied ultrasound radiomics to thyroid disease, demonstrating promise in differentiating benign and malignant thyroid nodules, predicting lymph node metastasis, and assessing molecular characteristics [13]. However, these imaging-based approaches are not specifically designed for pregnancy-related thyroid dysfunction and do not incorporate routine clinical variables available in early pregnancy. Consequently, machine-learning-based prediction models specifically developed for thyroid disease during pregnancy remain limited.

In this study, we used a retrospective maternal database and applied machine learning algorithms to identify risk factors and predict the risk of thyroid disease during pregnancy, with the aim of helping physicians to develop timely, personalized management strategies.

Material and Methods

Patients

This single-center retrospective study used anonymized data without identifiable information. Women who gave birth at Beijing Shijitan Hospital, Capital Medical University, between January 2020 and December 2024 were included. Clinical characteristics were recorded in the electronic medical record database and reviewed by obstetricians, obstetric nurses, and midwives. Patients with missing data for more than half of the selected feature variables were excluded. After data entry, quality management personnel in the obstetrics department reviewed the dataset to ensure accuracy. The study was approved by the Ethics Committee of Scientific Research at Beijing Shijitan Hospital, Capital Medical University (IIT2024-103-050); the informed consent requirement was waived due to the retrospective design. The study adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement [14].

Diagnostic Criteria for Thyroid Diseases

The primary outcome was a composite diagnosis of thyroid disease during pregnancy, including subclinical hypothyroidism, clinical hypothyroidism, and hyperthyroidism. Diagnoses were made by endocrinologists based on laboratory tests and clinical assessment, in accordance with the American Thyroid Association 2017 guidelines and institutional clinical pathways [9]. The operational definitions were as follows: (1) Clinical hypothyroidism: TSH concentration above the upper limit of the pregnancy-specific reference range (eg, >4.0 mIU/L in the first trimester at our institution) with decreased FT4 levels; (2) Subclinical hypothyroidism: elevated TSH (eg, >2.5 mIU/L in the first trimester) with normal FT4 levels; (3) Hyperthyroidism: suppressed TSH (eg, <0.1 mIU/L) with elevated FT4 and/or total triiodothyronine (T3), or a confirmed clinical diagnosis of Graves' disease. TPOAb status is an important risk factor but was not used as a mandatory diagnostic criterion for this composite outcome. This approach aligns with the study's aim of identifying biochemical thyroid dysfunction for initial screening.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Data Collection and Variables

Demographic information and pregnancy complications were collected from the electronic medical record database. A total of 18 predictive variables were included: age, height, pre-pregnancy weight, gravidity, parity, primiparity or multiparity, natural conception or in vitro fertilization and embryo transfer (IVF-ET), twin pregnancy, proteinuria during pregnancy, anemia during pregnancy, gestational diabetes mellitus (GDM), hypertensive disorders of pregnancy (HDP), scarred uterus, autoimmune disease (AID) during pregnancy, adverse pregnancy and childbirth history (APCD), thrombocytopenia during pregnancy, cardiovascular disease (CVD) during pregnancy, and respiratory disease during pregnancy. The outcome variable (binary) was thyroid disease during pregnancy (including subclinical hypothyroidism, clinical hypothyroidism, and hyperthyroidism). To establish a causal temporal relationship for prediction, all candidate predictor variables were required to be documented prior to the diagnosis of thyroid disease. For patients who developed thyroid disease, only complications recorded before the diagnosis date were included. For patients without thyroid disease, complications were assessed up to delivery. This approach ensured that the model supports true early risk prediction, using only information available at the time of clinical assessment.

Statistical Analysis

R (version 4.1.2) and RStudio (version 1.4.1106) were used for statistical analysis. Normally distributed continuous data are presented as mean±standard deviation (SD), and the independent-samples Student's t-test was used for comparisons between groups. Categorical data are presented as counts (n) and percentages (%), and the chi-square test was used to assess differences between groups. The combination of subclinical hypothyroidism, clinical hypothyroidism, and hyperthyroidism into a single composite outcome for model development was based on 3 considerations: (1) from a clinical management perspective, all 3 conditions require timely monitoring of thyroid function and appropriate intervention during pregnancy to reduce the risk of adverse maternal and fetal outcomes; (2) sample sizes for individual subtypes were insufficient to support robust subtype-specific models; and (3) the primary aim was to develop an early prescreening tool to identify pregnant women at risk for any thyroid dysfunction, rather than to predict specific subtypes. Univariate analyses were performed to assess associations between 18 independent variables (age, height, pre-pregnancy weight, gravidity, parity, primiparity or multiparity, IVF-ET, twin pregnancy, proteinuria, anemia, GDM, HDP, scarred uterus, AID, APCD, thrombocytopenia, CVD, and respiratory disease) and the dependent variable (thyroid disease). Notably, raw height and pre-pregnancy weight were used as separate features in machine learning models, rather than the derived body mass index (BMI). This approach allows greater flexibility for algorithms to capture complex, nonlinear relationships or interactions between these anthropometric measures that may be predictive of the outcome, which could be constrained by using a predefined index such as BMI. To prevent data leakage and ensure rigorous validation, a strict chronological split was applied, dividing the dataset into a training set (January 2020 to December 2023) and a hold-out test set (January 2024 to December 2024).

Feature variable selection was performed in the training set using the Boruta algorithm. After data normalization and standardization in the training set, 8 models were developed: logistic regression (LR), Bayesian approach (Bayes), k-nearest neighbors (KNN), support vector machine (SVM), neural network (NNET), classification and regression tree (CART), extreme gradient boosting (XGBoost), and random forest (RF). Model development incorporated a balanced sampling strategy (combining undersampling and oversampling), k-fold cross-validation (k=10), and hyperparameter optimization (HPO). Detailed cross-validated metrics for all candidate models are presented in **Table 1**. The best-performing model was selected and trained on the fully processed training set, then evaluated once on the untouched temporal test set to obtain an unbiased estimate of real-world performance. A feature importance ranking, univariate partial dependence plots, and breakdown profiles were generated. In both the training and test sets, the receiver operating characteristic (ROC) curve, precision-recall curve (PRC), calibration curve, and decision curve analysis (DCA) were assessed. Performance metrics, including accuracy (ACC), Matthews correlation coefficient (MCC), log loss, Brier score, area under the ROC curve (AUC-ROC), and area under the PRC (AUC-PRC), were calculated. The optimal threshold for the best model was determined by maximizing the Youden index (sensitivity+specificity-1) based on the ROC curve in the training set, then applied to the test set without further adjustment to avoid overfitting and provide an unbiased estimate of real-world performance. A conventional multivariable LR model incorporating age, BMI, parity, and AID was constructed as a benchmark for comparison with the best model [5,9].

A sample size calculation was performed using the pmsampsize function in RStudio. For the binary prediction model, a C-statistic of 0.9 was assumed, with 18 predictor variables and an incidence rate of thyroid disease during pregnancy of 0.5%. This incidence rate was conservatively chosen based on the reported incidence of clinical hypothyroidism (0.3% to 0.5%) in the general pregnant population from prior literature [2,3], which represented the lower bound of thyroid disease subtypes included in our composite outcome. Although the observed composite outcome rate in our hospital-based cohort was higher (12.07% in the training set), the use of a lower incidence rate provided a more conservative sample size estimate. The calculation yielded a Cox-Snell R^2 of 0.1278, indicating that

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 1. Cross-validated performance metrics for all models.

nr task_id	learner_id	resampling_id	iters	classif.ce
1: 1 train	LR	repeated_cv	100	0.4367236
1: 1 train	Bayes	repeated_cv	100	0.4642220
1: 1 train	KKNN	repeated_cv	100	0.1699081
1: 1 train	SVM	repeated_cv	100	0.4135028
1: 1 train	NNET	repeated_cv	100	0.4201979
1: 1 train	CART	repeated_cv	100	0.4253256
1: 1 train	XGBoost	repeated_cv	100	0.3768267
1: 1 train	RF	repeated_cv	100	0.1416724

Abbreviations: Bayes, Bayesian approach; CART, classification and regression tree; KNN, k-nearest neighbors; LR, logistic regression; NNET, neural network; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

Table 2. Patient clinical characteristics.

Characteristics	Training set (n=4466)			Test set (n=995)		
	Normal (n=3927)	Thyroid disease (n=539)	*P* value	Normal (n=894)	Thyroid disease (n=101)	*P* value
Age, mean (SD), y	31.83 (3.83)	32.02 (3.80)	0.264	32.34 (3.97)	32.53 (4.05)	0.641
Height, mean (SD), cm	162.71 (5.05)	162.33 (5.12)	0.103	162.69 (5.18)	162.97 (5.26)	0.600
Pre-pregnancy weight, mean (SD), kg	58.17 (9.35)	59.22 (10.08)	0.015	58.59 (9.24)	59.96 (9.34)	0.159
BMI, mean (SD), kg/m^2	21.95 (3.24)	22.44 (3.44)	0.001	22.12 (3.24)	22.54 (3.06)	0.224
Gravidity, mean (SD)	1.77 (0.99)	1.84 (0.97)	0.113	1.64 (0.94)	1.78 (0.99)	0.161
Parity, mean (SD)	1.35 (0.51)	1.30 (0.47)	0.056	1.30 (0.51)	1.35 (0.56)	0.345
PM, No. (%)			0.147			0.512
Primipara	2627 (66.9)	378 (70.1)		652 (72.9)	70 (69.3)	
Multipara	1300 (33.1)	161 (29.9)		242 (27.1)	31 (30.7)	
IVF-ET, No. (%)	109 (2.8)	14 (2.6)	0.923	10 (1.1)	0 (0.0)	0.588
Twins, No. (%)			0.938			00.172
Single	3885 (98.9)	534 (99.1)		869 (97.2)	101 (100.0)	
Twins	42 (1.1)	5 (0.9)		25 (2.8)	0 (0.0)	
Proteinuria, No. (%)	55 (1.4)	8 (1.5)	1	13 (1.5)	2 (2.0)	1
Anemia, No. (%)			0.64			0.529
No	2715 (69.1)	375 (69.6)		844 (94.4)	98 (97.0)	
Mild	997 (25.4)	141 (26.2)		49 (5.5)	3 (3.0)	
Moderate	212 (5.4)	23 (4.3)		1 (0.1)	0 (0.0)	

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 2 continued. Patient clinical characteristics.

Characteristics	Training set (n=4466)			Test set (n=995)		
	Normal (n=3927)	Thyroid disease (n=539)	*P* value	Normal (n=894)	Thyroid disease (n=101)	*P* value
Severe	3 (0.1)	0 (0.0)		0 (0.0)	0 (0.0)	
GDM, No. (%)	626 (15.9)	104 (19.3)	0.056	174 (19.5)	21 (20.8)	0.852
HDP, No. (%)			0.3			0.307
No	3668 (93.4)	496 (92.0)		781 (87.4)	88 (87.1)	
Chronic	34 (0.9)	2 (0.4)		7 (0.8)	3 (3.0)	
Gestational	78 (2.0)	16 (3.0)		44 (4.9)	4 (4.0)	
Chronic preeclampsia	30 (0.8)	3 (0.6)		0 (0.0)	0 (0.0)	
Preeclampsia	58 (1.5)	12 (2.2)		25 (2.8)	3 (3.0)	
Eclampsia	59 (1.5)	10 (1.9)		37 (4.1)	3 (3.0)	
Scarred uterus, No. (%)	281 (7.2)	37 (6.9)	0.061	58 (6.5)	7 (6.9)	1
AID, No. (%)	16 (0.4)	4 (0.7)	0.455	3 (0.3)	0 (0.0)	1
APCD, No. (%)	84 (2.1)	19 (3.5)	0.063	0 (0.0)	0 (0.0)	NA
Thrombocytopenia, No. (%)	62 (1.6)	9 (1.7)	1	15 (1.7)	2 (2.0)	1
CVD, No. (%)	87 (2.2)	15 (2.8)	0.501	3 (0.3)	1 (1.0)	0.876
Respiratory disease, No. (%)	13 (0.3)	5 (0.9)	0.092	6 (0.7)	1 (1.0)	1

Note: BMI was calculated as pre-pregnancy weight in kilograms divided by height in meters squared. For age, height, pre-pregnancy weight, BMI, gravidity, and parity, the independent-samples Student's t-test was used to compare differences between groups. For PM, IVF-ET, twins, proteinuria, anemia, GDM, HDP, scarred uterus, AID, APCD, thrombocytopenia, CVD, and respiratory disease, the χ^2 test or Fisher's exact test was used for group comparisons. All *P* values are presented for descriptive and exploratory purposes only; they were derived from univariate analyses. Variable selection for the prediction models was based on the Boruta algorithm and machine learning procedures, rather than univariate significance testing. Abbreviations: AID, autoimmune disease; APCD, adverse pregnancy and childbirth history; BMI, body mass index; CVD, cardiovascular disease; GDM, gestational diabetes mellitus; HDP, hypertensive disorders of pregnancy; IVF-ET, in vitro fertilization and embryo transfer; PM, primipara or multipara; SD, standard deviation.

at least 1176 samples were required to achieve sufficient statistical power. This requirement was met by the 4466 samples available in the training set [15-19].

Results

Patient Characteristics

Among the 5636 patients initially reviewed, 175 were excluded due to more than 50% missing data in the selected feature variables. This retrospective study ultimately included 5461 patients: 4466 in the training set and 995 in the test set. Thyroid disease during pregnancy occurred in 539 cases (12.07%) in the training set and 101 cases (10.15%) in the test set. In the training set, pre-pregnancy weight (59.22±10.08 kg) and BMI (22.44±13.44 kg/m^2) were significantly higher in patients with thyroid disease during pregnancy than in those without (58.17±9.35 kg and 21.95±3.24 kg/m^2, respectively) (P<0.05). Detailed patient characteristics are presented in **Table 2**. A flowchart illustrating patient selection, data segmentation, model development, testing, and interpretation is shown in **Figure 1**.

Model Development and Selection

In the training set, 9 of the 18 independent variables were identified as feature variables through the Boruta algorithm (age, height, pre-pregnancy weight, gravidity, parity, primiparity or multiparity, HDP, scarred uterus, and AID) and were used to develop 8 models (**Figure 2**). Among candidate models, the RF

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

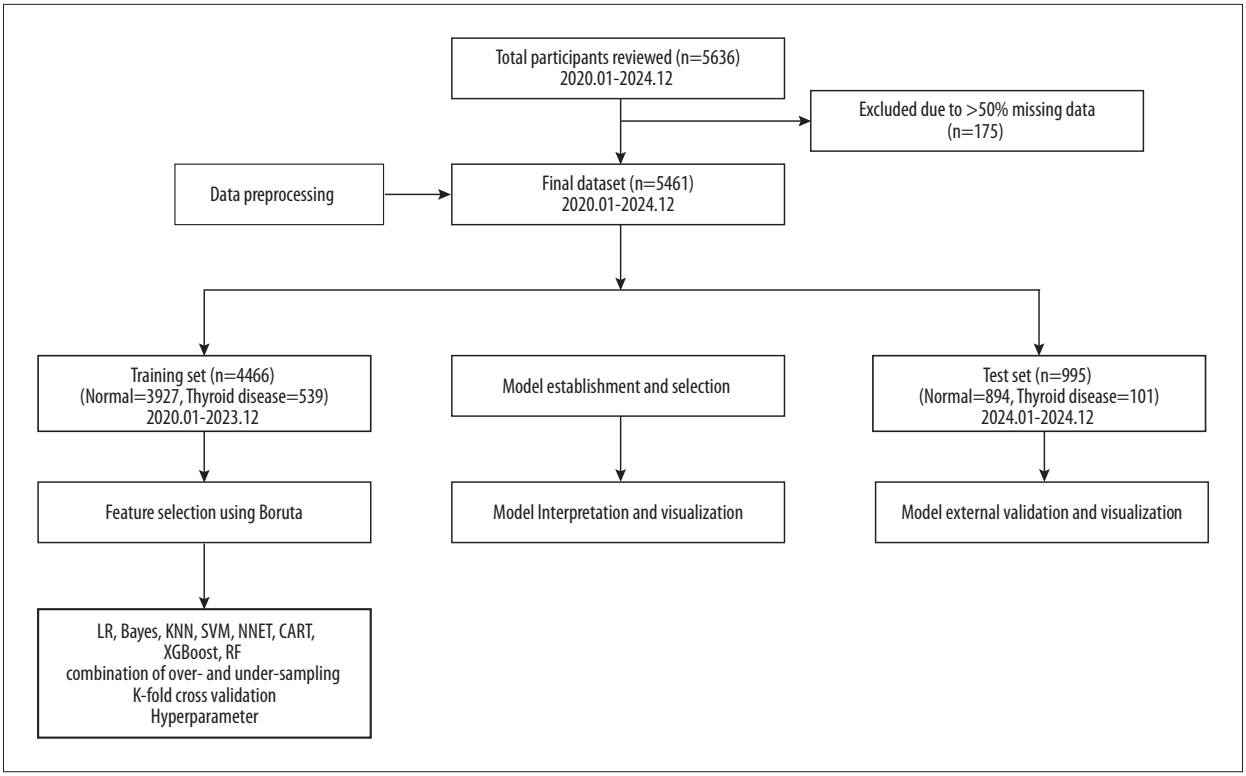

Figure 1. Flowchart of patient selection, data segmentation, and model development. The flowchart illustrates the processes of data segmentation, feature selection, model development and interpretation, external validation, and visualization. Abbreviations: Bayes, Bayesian approach; CART, classification and regression tree; KNN, k-nearest neighbors; LR, logistic regression; NNET, neural network; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

model demonstrated the best performance in terms of AUC-ROC and AUC-PRC (**Figure 3**). Optimized hyperparameters included num.threads=1, mtry=1, min.node.size=1, and num.trees=982.

ROC, PRC, Calibration Curve, DCA, and Confusion Matrix Parameters

For prediction of thyroid disease during pregnancy using the RF model, analyses included ROC, PRC, calibration curve, and DCA. In the training and test sets, respectively, the ACC was 0.98387819 and 0.99597990, MCC was 0.96794139 and 0.97781292, log loss was 0.12670703 and 0.09442025, Brier score was 0.02495798 and 0.01921069, AUC-ROC was 0.99877170 and 0.99991140, and AUC-PRC was 0.99864486 and 0.99922572. The optimal threshold for the RF model, determined from the ROC curve, was 0.518. The ROC, PRC, calibration curve, and DCA for the test set are shown in **Figure 4**. For benchmark comparison, the LR model based on conventional risk factors is compared with the RF model in **Table 3**.

Feature Importance Ranking and Univariate Partial Dependence

Based on the RF model, feature importance (**Figure 5**) and univariate partial dependence (**Figure 6**) for the 9 variables were analyzed. **Figure 5** illustrates the relative contribution of each feature to thyroid disease during pregnancy. **Figure 6** demonstrates the effect of each feature on the outcome and the corresponding trend as the feature value changes.

Breakdown Profile for a Randomly Selected Single Sample

The breakdown profile illustrates the contribution of each variable to the prediction for a single sample (**Figure 7**). Red and blue bars represent the positive and negative contributions of each variable, respectively; the predicted value equals the sum of contributions from all features. The RF model predicted a probability of 0.583 for this sample, exceeding the threshold of 0.518. Therefore, the model classified this patient as likely to develop thyroid disease. Consistent with this prediction, the patient was diagnosed with thyroid disease during pregnancy.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

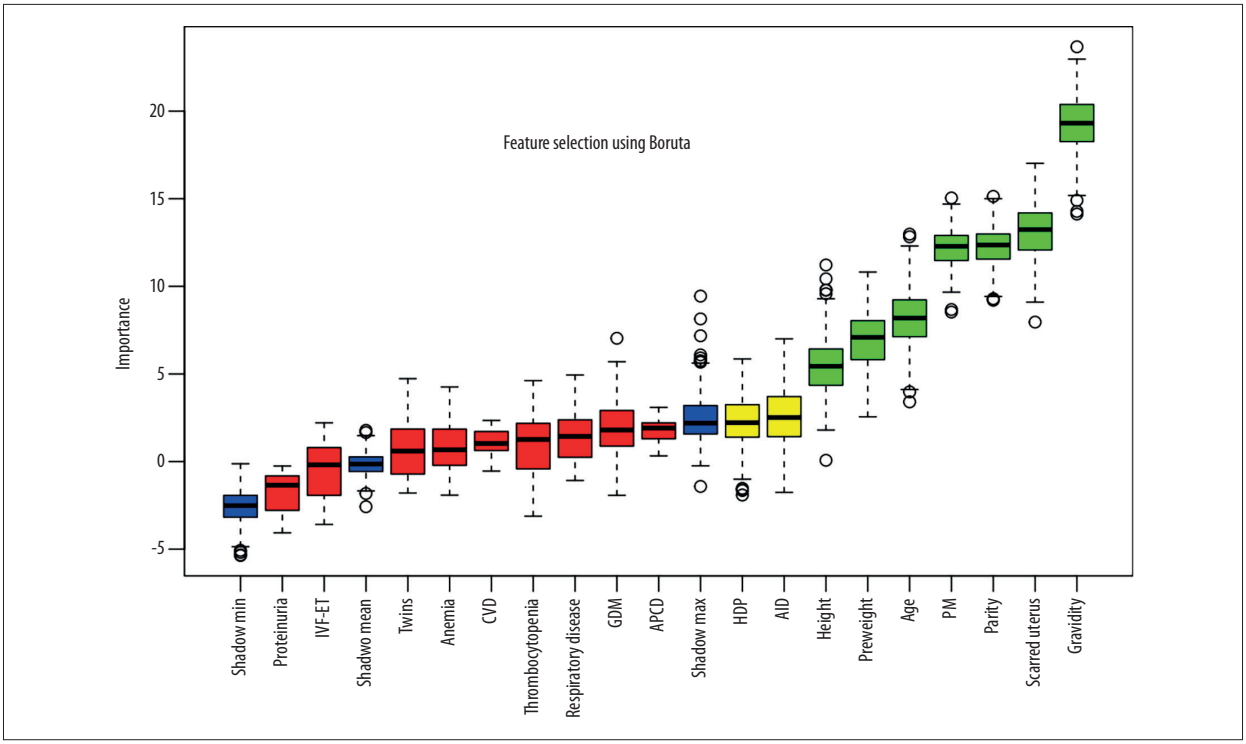

Figure 2. Candidate variable screening diagram based on Boruta. The figure illustrates the results of feature selection using the Boruta algorithm. Feature variables are displayed along the x-axis, and their importance scores are shown on the y-axis. Green boxes represent features confirmed as important ("Confirmed"); red boxes denote features rejected as unimportant ("Rejected"); and blue boxes indicate the minimum, median, and maximum importance of random shadow features used for comparison ("Shadow"). Abbreviations: AID, autoimmune disease; APCD, adverse pregnancy and childbirth history; CVD, cardiovascular disease; GDM, gestational diabetes mellitus; HDP, hypertensive disorders of pregnancy; IVF-ET, in vitro fertilization and embryo transfer; PM, primipara or multipara.

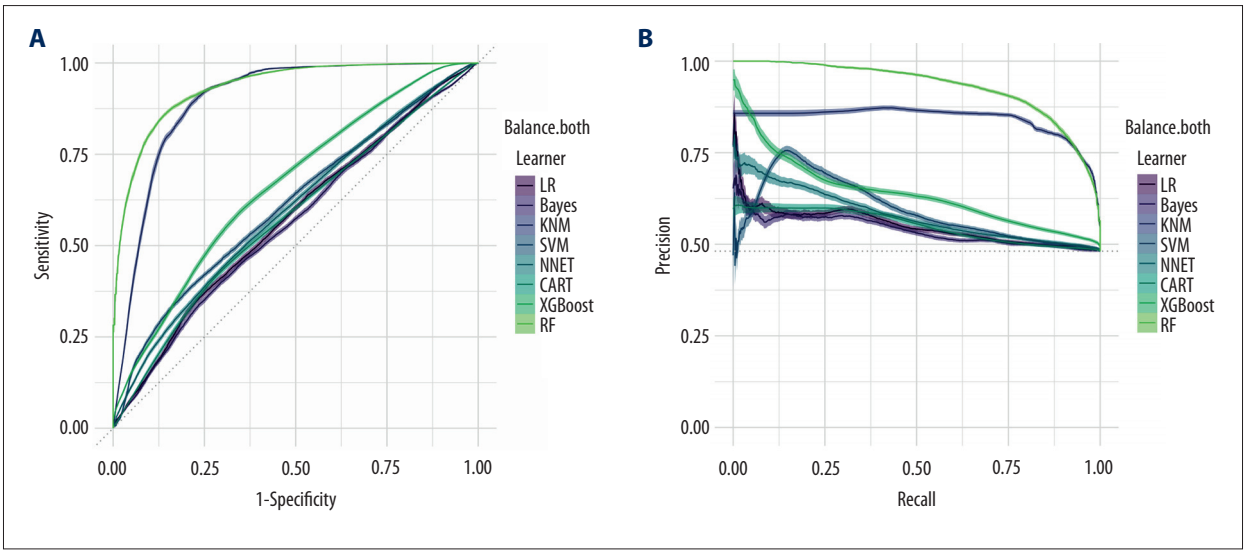

Figure 3. Model comparison and selection. The figure shows the performance of 8 models (LR, Bayes, KNN, SVM, NNET, CART, XGBoost, and RF) using ROC and PRC to predict thyroid disease during pregnancy. (**A**) ROC curves of the 8 models; (**B**) PRCs of the 8 models. Abbreviations: Bayes, Bayesian approach; CART, classification and regression tree; KNN, k-nearest neighbors; LR, logistic regression; NNET, neural network; PRC, precision-recall curve; RF, random forest; ROC, receiver operating characteristic curve; SVM, support vector machine; XGBoost, extreme gradient boosting.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

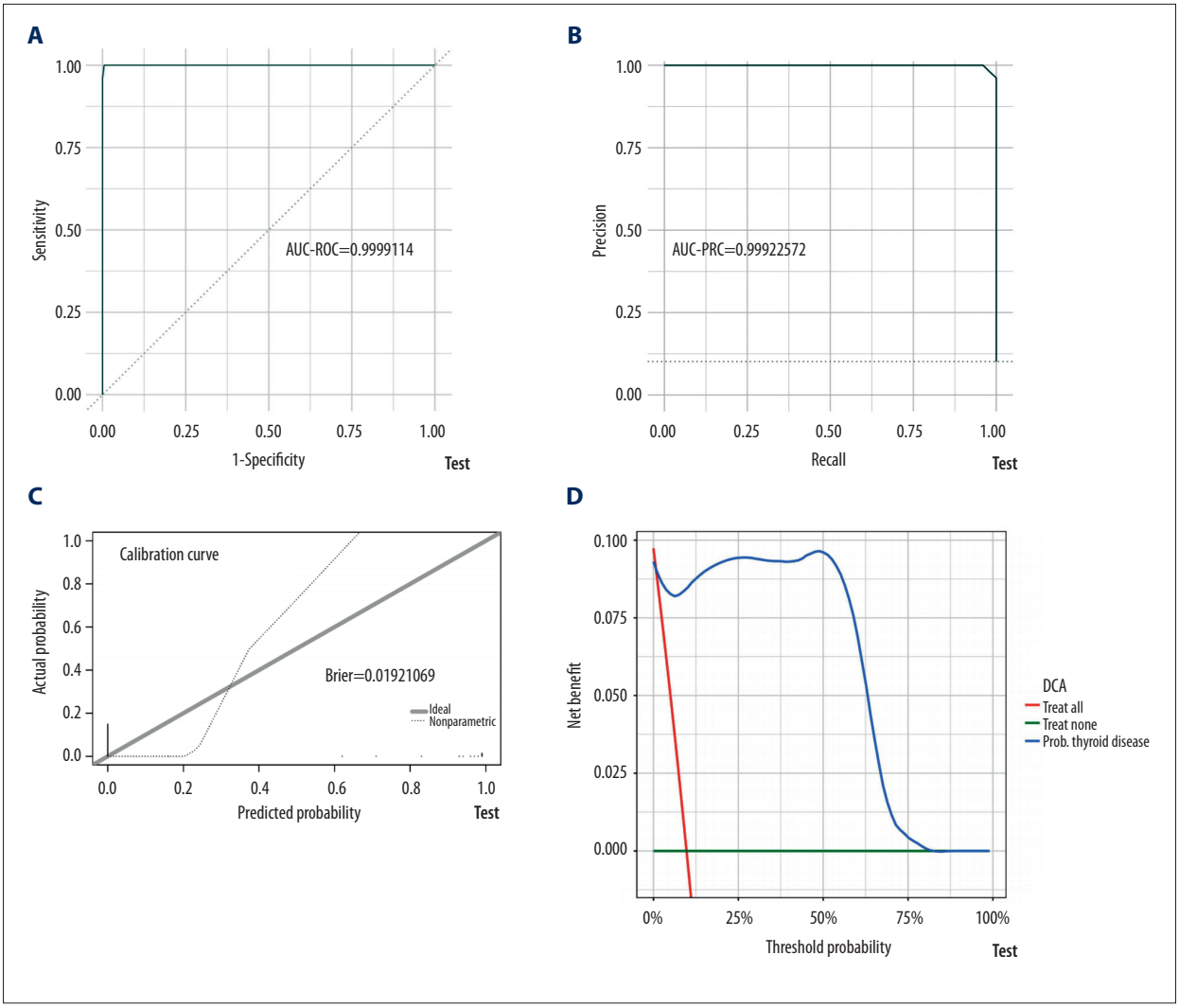

Figure 4. ROC, PRC, calibration curve, and DCA in the test set. (**A**) ROC curve showing the relationship between sensitivity and specificity of the RF model. (**B**) PRC illustrating the trade-off between precision and recall. (**C**) Calibration curve demonstrating agreement between predicted probabilities and observed incidence. (**D**) DCA assessing the net benefit of the RF model across different threshold probabilities. Abbreviations: AUC-PRC, area under the PRC; AUC-ROC, area under the ROC curve; DCA, decision curve analysis; PRC, precision-recall curve; RF, random forest; ROC, receiver operating characteristic curve.

Discussion

Thyroid disease during pregnancy is challenging to predict and manage clinically due to complex risk factors. The development of a precise assessment system is essential to facilitate early detection and intervention. Our results indicate that the machine-learning-based RF model outperforms other models in predicting thyroid disease during pregnancy.

As a core concept in machine learning, feature variables have a critical impact on model performance. Their importance includes the following aspects: (1) foundation of model performance—high-quality features can substantially enhance model performance; (2) transformation of data—feature variables convert raw data into a format that algorithms can process; and (3) influence on complexity—the number of features directly affects model complexity [20,21]. Therefore, effective feature engineering often contributes more to model performance than the selection of increasingly complex algorithms, making it a stage that warrants substantial time investment in machine learning projects. Based on these considerations, the present study utilized the embedded Boruta feature selection algorithm. Boruta is an RF-based method that identifies which feature variables in a dataset genuinely contribute to the prediction of target outcomes. It is more stringent and reliable than conventional feature importance methods, effectively distinguishing

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 3. Benchmark comparison of model performance.

Performance metric	LR model with conventional risk factors		RF model	
	Training set	Test set	Training set	Test set
AUC-ROC	0.55608314	0.55041310	0.99877170	0.99991140
AUC-PRC	0.53579933	0.12096839	0.99864486	0.99922572
ACC	0.55127631	0.89849246	0.98387819	0.99597990
Brier score	0.24692233	0.09098832	0.02495798	0.01921069
Log loss	0.68693139	0.32699989	0.12670703	0.09442025
MCC	0.09550202	0.00	0.96794139	0.97781292

Abbreviations: ACC, accuracy; AUC-ROC, area under the receiver operating characteristic curve; AUC-PRC, area under the precision-recall curve; LR, logistic regression; MCC, Matthews correlation coefficient; RF, random forest.

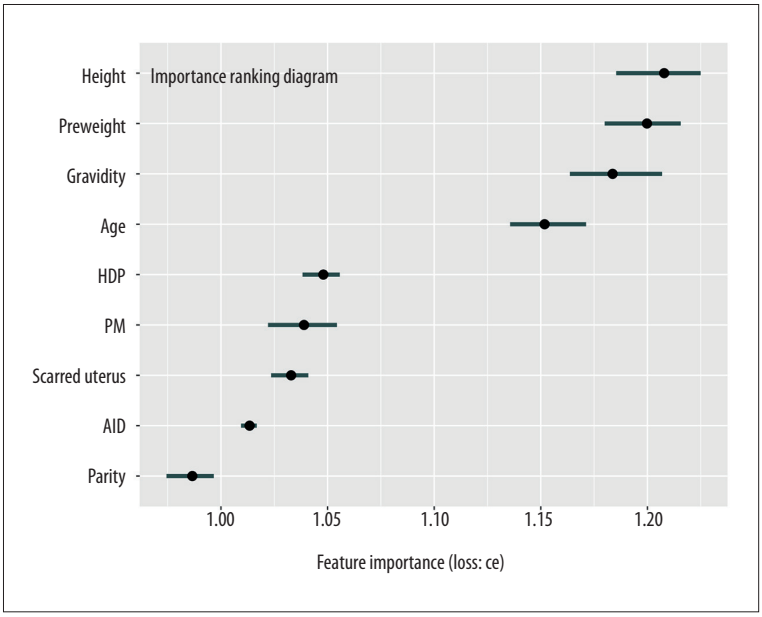

Figure 5. Importance ranking of feature variables. The diagram illustrates the relative contribution of each feature variable to the prediction of thyroid disease during pregnancy. Abbreviations: AID, autoimmune disease; HDP, hypertensive disorders of pregnancy; PM, primipara or multipara; Preweight, pre-pregnancy weight.

relevant features from noise. This makes it a more robust approach than single-instance feature importance assessments, particularly for complex datasets requiring high-confidence feature selection [22]. The core principle of Boruta is that a feature is considered truly important only if it performs better than its randomized counterpart (shadow feature). It offers several advantages: (1) comprehensiveness—it adjusts for interdependencies among features; (2) reliability—it reduces misclassification through multiple iterations and statistical testing; (3) nonparametric nature—it does not require predefined importance thresholds; and (4) intuitive interpretation—the results are straightforward to understand [20,23].

In the present study, the Boruta feature selection process identified 9 variables associated with thyroid disease during pregnancy: age, height, pre-pregnancy weight, gravidity, parity, primiparity, HDP, scarred uterus, and AID. Notably, the partial dependence and breakdown analyses indicated that height, pre-pregnancy weight, age, and gravidity contributed most to the prediction of thyroid disease during pregnancy. A prospective cohort study by Yang et al revealed that increased pre-pregnancy weight was associated with elevated maternal TSH levels and thyroid dysfunction, consistent with our findings [24]. A meta-analysis of individual participant data showed that maternal age over 30 years and higher maternal BMI were associated with abnormal thyroid function tests, suggesting value in guiding thyroid disease screening during pregnancy [5]. Additionally, a recent systematic review concluded that advanced maternal age is associated with increased risks of subclinical hypothyroidism, isolated hypothyroxinemia, and abnormal thyroid function tests [4].

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

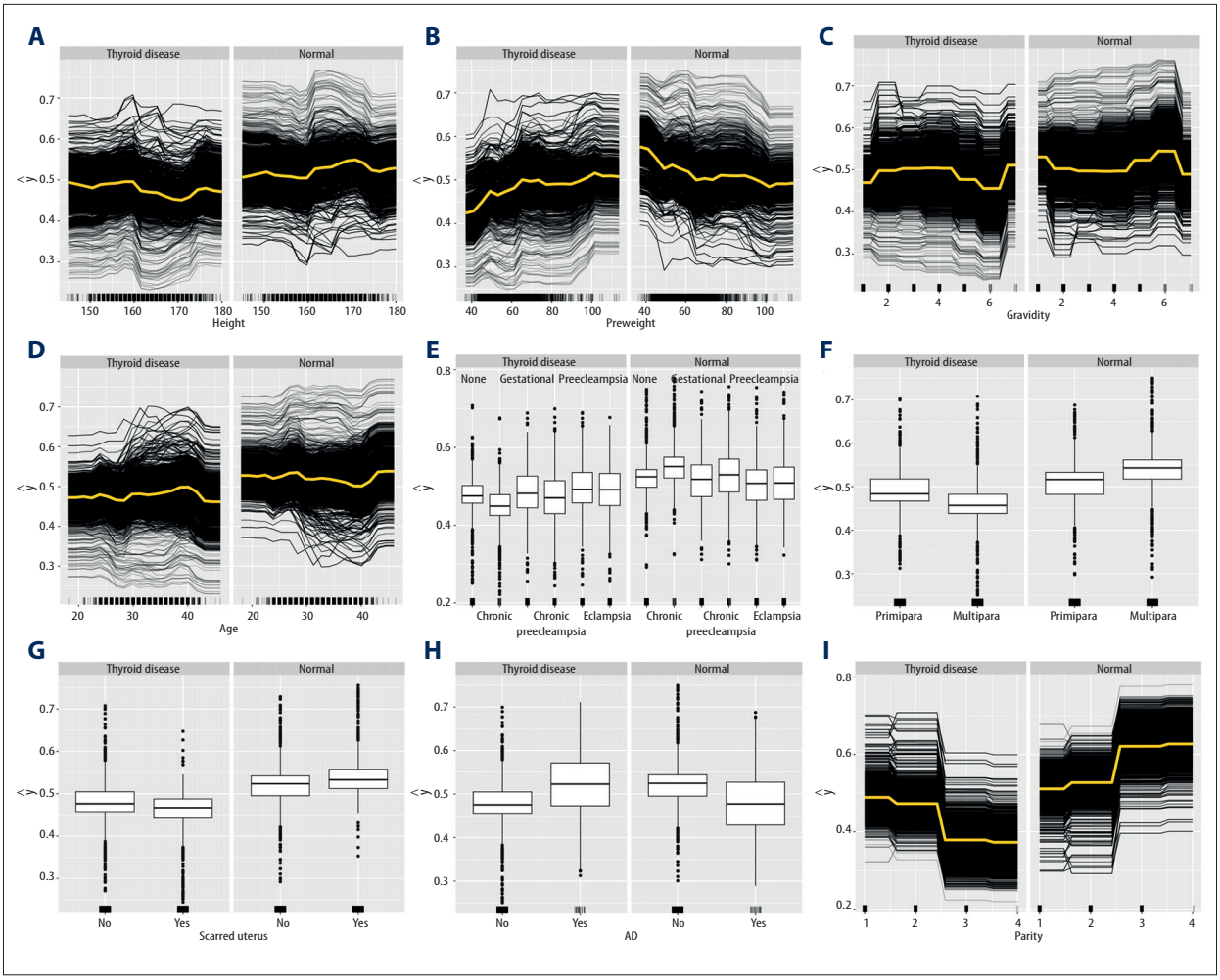

Figure 6. Univariate partial dependence profile. Univariate partial dependence profiles for thyroid disease during pregnancy, illustrating the effect of each feature variable on the RF model predictions. Each subfigure represents the average predicted response across a range of values for a given feature: (**A**) height; (**B**) pre-pregnancy weight; (**C**) gravidity; (**D**) age; (**E**) HDP; (**F**) PM; (**G**) scarred uterus; (**H**) AID; and (**I**) parity. Abbreviations: AID, autoimmune disease; HDP, hypertensive disorders of pregnancy; PM, primipara or multipara; Preweight, pre-pregnancy weight; RF, random forest.

The interrelationship among high-risk pregnancies, adverse neonatal outcomes, and thyroid disease during pregnancy should not be overlooked. As a common complication in high-risk pregnancies, coagulation dysfunction can be assessed using laboratory markers such as D-dimer and soluble fibrin monomer complex. A recent observational study demonstrated that soluble fibrin monomer complex concentrations remain relatively stable during pregnancy and are less affected by gestational age than D-dimer, making soluble fibrin monomer complex a potentially more reliable marker of thrombosis risk in pregnant women [25]. This finding suggests the potential value of coagulation dysfunction, as reflected by different laboratory markers, in predicting thyroid disease during pregnancy. In high-risk pregnancies, Doppler ultrasound parameters—including the resistance index and pulsatility index of the umbilical artery and the cerebroplacental ratio—have shown utility in predicting adverse neonatal outcomes. In particular, an abnormal antenatal umbilical coiling index detected by Doppler ultrasound is significantly associated with increased risks of fetal growth restriction and a low 5-minute Apgar score [26]. Therefore, incorporation of Doppler-ultrasound-based hemodynamic assessment to monitor neonatal outcomes may be particularly important in high-risk pregnancies, especially those complicated by thyroid disease. However, the current prediction model is designed as a prescreening tool solely based on routine clinical variables. Future studies integrating coagulation markers and Doppler ultrasound parameters into this model may further improve the identification of pregnancies at risk for both thyroid dysfunction and adverse neonatal outcomes, enabling more comprehensive risk stratification and personalized management.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

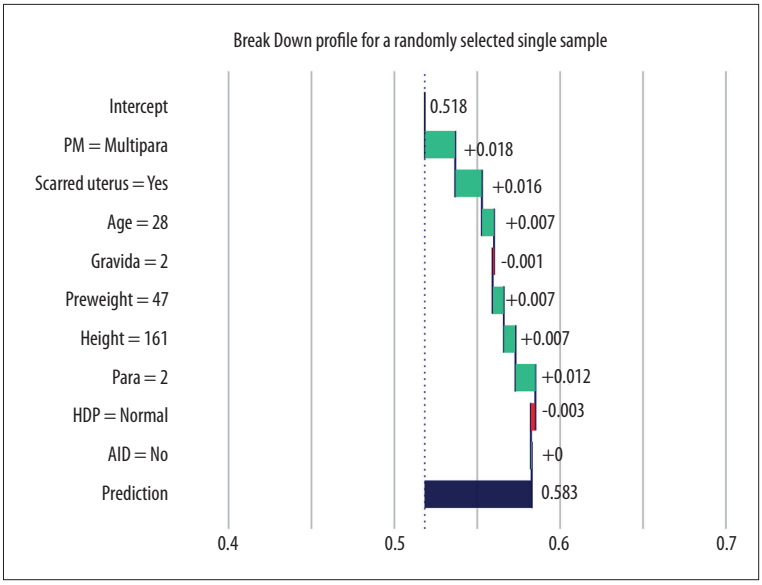

Figure 7. Break down profile
Abbreviations: PM, primipara or multipara; Preweight, pre-pregnancy weight; HDP, hypertensive disorders of pregnancy; AID, autoimmune disease. Note: The profile illustrates the contribution of individual features to the final prediction result for a single randomly selected sample. The red and blue bars in the figure represent the positive and negative impacts of each variable on the prediction. The predicted value equals the sum of contributions from each feature variable.

Multiple studies have explored the relationship between gravidity and thyroid disease during pregnancy, although the evidence remains limited. A cross-sectional study involving 54 586 singleton pregnancies identified an association between increased gravidity and the occurrence of isolated maternal hypothyroxinemia [27]. Similarly, a recent prospective cohort study indicated that gravidity ≥ 2 was associated with TPOAb positivity and could enhance the exposure-response relationship between particulate matter and TPOAb positivity [28]. According to multivariable LR analysis, Liu et al concluded that increased gravidity was an independent risk factor for adverse pregnancy outcomes among patients with gestational hypothyroidism [29]. Given the substantial contribution of gravidity to thyroid disease during pregnancy observed in the present study, further research is warranted to examine its role across different thyroid disease subtypes.

Although thyroid disease during pregnancy was defined as a composite outcome in the present study, subclinical hypothyroidism, clinical hypothyroidism, and hyperthyroidism differ regarding pathophysiology, risk factors, and clinical outcomes. Consequently, the predictive performance of the model may vary across subtypes. Hu et al developed separate machine learning models for hyperthyroidism and hypothyroidism using routine laboratory tests; they reported AUC-ROC values of 93.8% and 90.9%, respectively, suggesting that different subtypes may have distinct feature associations and predictive accuracy [30]. Our study did not clarify whether the model performs equally well across subtypes. Certain variables (eg, AID) may be more predictive of hypothyroidism, whereas others (eg, multiple gestation) may be more relevant to hyperthyroidism. Future studies with larger subtype-specific sample sizes are needed to develop and validate models tailored to each subtype. Moreover, the use of a composite outcome has inherent clinical limitations because it does not distinguish among specific subtypes or guide subtype-specific management. Patients identified as high risk by the model should undergo confirmatory biochemical testing for accurate diagnosis and appropriate treatment. Clinicians should also note that high model performance was achieved using a single-center dataset with temporal validation; external validation in diverse populations is required to confirm generalizability across subtypes and clinical settings.

Based on these considerations, our model should undergo prospective external validation in multicenter cohorts before clinical implementation. After validation, it could be utilized in early pregnancy (eg, at the first antenatal visit) and integrated into electronic medical records to automatically identify patients at high risk for thyroid disease during pregnancy. Because clinical status and risk may evolve over time, the model should not be limited to a single time point. It is also recommended that the model be reapplied when new clinical variables emerge (eg, new-onset HDP), enabling dynamic risk stratification.

MLR3 is a modern machine learning framework in the R programming language. Through its modular design and extensive ecosystem of extensions, it provides a powerful and flexible toolkit for machine learning tasks. It offers a unified interface and diverse functionalities to support multiple stages of the machine learning workflow, including model evaluation, flexible resampling strategies, model comparison, optimization methods, data preprocessing, feature selection and transformation, and visualization. Accordingly, the present study used multimodel selection within the MLR3 framework to evaluate the performance of various algorithms under data perturbations and to identify the optimal model for the dataset. This approach helps avoid suboptimal solutions associated with reliance on a

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

single model and ensures a comprehensive process from data exploration to model development [31]. Furthermore, k-fold cross-validation (k=10) and HPO were utilized for model selection. Cross-validation is a standard method for evaluating and selecting machine learning models. It involves repeated partitioning of the dataset into training and validation subsets, using a rotational validation approach to obtain more reliable performance estimates, which provides a robust basis for model selection and optimization [32,33]. HPO further improves model performance and generalizability [34]. Thus, the combination of k-fold cross-validation and HPO facilitated development of a model well suited to the dataset.

We note that the highly imbalanced class distribution in our dataset may adversely affect the performance of standard algorithms. Therefore, efforts to address class imbalance are essential for improved model performance [35]. In the present study, a balanced sampling strategy combining undersampling and oversampling was applied to the training set for model development, thereby reducing training bias and improving the decision boundary [36].

External validation is a critical component of machine-learning-based model development, which involves evaluation of model performance using a dataset entirely independent of the training process to provide an unbiased estimate of real-world performance. Here, we conducted temporal external validation by using later data to validate models developed on earlier data, thus assessing robustness to temporal drift. Additionally, the test set was not resampled during validation, potentially yielding results more reflective of real-world conditions [37].

After model development, performance evaluation is equally important. Various metrics capture different aspects of performance. Our study used the following: (1) ACC, to reflect overall predictive accuracy; (2) MCC, as a comprehensive metric incorporating all elements of the confusion matrix; (3) log loss, to penalize incorrect probability estimates; (4) Brier score, to assess both discrimination and calibration; (5) AUC-ROC, to evaluate overall discriminative ability; and (6) AUC-PRC, to assess performance in identifying positive (minority) cases. This multi-metric evaluation approach provided a comprehensive assessment of model performance and avoided reliance on any single metric.

In the present study, the optimal threshold for the RF model was identified as 0.518 based on the maximum Youden index from the ROC curve. However, this threshold can be adjusted according to the clinical context. For example, lowering the threshold increases sensitivity for population-based screening, identifying more high-risk individuals at the cost of increased false positives. Conversely, raising the threshold improves specificity for confirmatory testing. Any modification of the threshold should be validated in an independent cohort before clinical implementation.

Regarding interpretation of machine learning models, feature importance ranking, partial dependence plots, and breakdown profiles are key tools for understanding model behavior, feature contributions, and prediction logic. These methods provide complementary perspectives on the model's decision-making process, supporting validation and identifying potential issues. In the present study, the feature importance ranking illustrates the overall contribution of each variable, enabling rapid identification of the most predictive features and providing global interpretation of the model [38]. The partial dependence plot demonstrates the monotonic or nonlinear relationship between an individual feature and the predicted outcome, as well as potential threshold effects [39]. In contrast, the breakdown profile focuses on individual predictions by decomposing the contribution of each feature, thus revealing how each variable influences the outcome for a given sample [38]. Collectively, these methods provide a multidimensional framework for interpreting model decisions, improving the transparency of nonparametric "black box" models, and enhancing model credibility and practical applicability [38-40].

A key strength of this study is its strict adherence to temporal precedence. By ensuring that all predictor variables were collected before the diagnosis of thyroid disease, the model functions as a true predictive tool rather than a retrospective classifier. This methodological rigor suggests that the identified associations—such as those between early-pregnancy BMI or prior AID and subsequent thyroid dysfunction—reflect potential causal pathways relevant to early screening. It also indicates that the model's performance estimates more closely approximate its real-world clinical utility, where only baseline and early-pregnancy data are available for decision-making. Furthermore, this study directly evaluated clinical utility by benchmarking the machine learning model against a conventional guideline-based approach. The substantial performance difference highlights the incremental value of the machine learning approach: the RF model achieved an AUC-ROC of 0.999, whereas the LR model achieved 0.550. These findings suggest that complex, nonlinear interactions among routine clinical variables, effectively captured by the RF algorithm, can provide more accurate risk stratification than conventional linear models based on a limited set of predefined factors.

For clinical decision-making, we recommend that patients identified as high risk (predicted probability >0.518) undergo confirmatory biochemical testing (TSH, FT4, and TPOAb). In contrast, low-risk patients may avoid unnecessary testing, thereby reducing costs and anxiety. The model is intended as a decision-support tool and not a replacement for clinical judgment.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Final decisions should integrate model outputs with individual patient characteristics and physician expertise.

This study has some limitations. First, despite the inclusion of multiple clinical variables, key factors emphasized in clinical guidelines were not available. These include thyroid-specific biomarkers (TSH, FT4, TPOAb, TgAb, and TRAb), iodine status (a gold-standard predictor of thyroid dysfunction), personal or family history of thyroid disease, and specific autoimmune conditions such as type 1 diabetes. These biomarkers and factors represent standard diagnostic criteria. Due to limitations of the retrospective electronic medical record database, the absence of such markers and factors indicates that the model cannot replace biochemical testing. Instead, it should be considered a prescreening tool that uses readily available clinical data to identify high-risk individuals for prioritized laboratory evaluation. Second, nonclinical patient factors, such as psychological and socioeconomic variables that may also influence thyroid disease during pregnancy, were not included [41,42]. Third, imaging data were not collected. Specifically, structural imaging data such as thyroid ultrasound were not routinely available for screening in this cohort. Although initial screening for thyroid disease during pregnancy primarily relies on serum biomarkers rather than imaging, incorporation of additional data modalities may improve predictive performance. Fourth, although the modeling strategy was designed to minimize overfitting, the near-perfect performance metrics should be interpreted with caution. These results likely reflect strong performance within this single-center cohort and the effectiveness of the temporal validation approach; however, the potential for residual overfitting cannot be excluded. Finally, although temporal external validation improves real-world relevance, the single-center design may limit generalizability. Therefore, future multicenter studies incorporating TSH/FT4 data and independent datasets collected across different times and locations are needed to further evaluate the model's external validity and generalizability.

Conclusions

We developed 8 machine-learning-based models, and the RF model demonstrated superior performance in predicting thyroid disease during pregnancy, supporting its potential as a reliable clinical tool. These findings enhance the potential for early identification and intervention in thyroid disease during pregnancy; they provide clinicians with a valuable tool to improve risk assessment and decision-making.

Declaration of Figures' Authenticity

All figures submitted have been created by the authors who confirm that the images are original with no duplication and have not been previously published in whole or in part.

References:

1. Yap YW, Onyekwelu E, Alam U. Thyroid disease in pregnancy. Clin Med (Lond). 2023;23(2):125-28
2. Dickens LT, Cifu AS, Cohen RN. Diagnosis and management of thyroid disease during pregnancy and the postpartum period. JAMA. 2019;321(19):1928-29
3. Stagnaro-Green A, Pearce E. Thyroid disorders in pregnancy. Nat Rev Endocrinol. 2012;8(11):650-58
4. Liu Y, Osinga JAJ, Maraka S, et al. Risk factors for thyroid function test abnormalities during pregnancy: A systematic review of the literature to validate current risk factors and identify novel ones. Thyroid. 2025;35(5):553-75
5. Osinga JAJ, Liu Y, Männistö T, et al. Risk factors for thyroid dysfunction in pregnancy: An individual participant data meta-analysis. Thyroid. 2024;34(5):646-58
6. Korevaar TI, de Rijke YB, Chaker L, et al. Stimulation of thyroid function by human chorionic gonadotropin during pregnancy: A risk factor for thyroid disease and a mechanism for known risk factors. Thyroid. 2017;27(3):440-50
7. Casey B, de Veciana M. Thyroid screening in pregnancy. Am J Obstet Gynecol. 2014;211(4):351-53.e1
8. Singh P, Boelaert K. Controversies in thyroid disease management in pregnancy. Clin Med (Lond). 2025;25(1):100287
9. Alexander EK, Pearce EN, Brent GA, et al. 2017 guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. Thyroid. 2017;27(3):315-89
10. Lee SY, Pearce EN. Assessment and treatment of thyroid disorders in pregnancy and the postpartum period. Nat Rev Endocrinol. 2022;18(3):158-71
11. Toro-Tobon D, Loor-Torres R, Duran M, et al. Artificial intelligence in thyroidology: A narrative review of the current applications, associated challenges, and future directions. Thyroid. 2023;33(8):903-17
12. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58
13. Lu WW, Zhang D, Ni XJ. A review of the role of ultrasound radiomics and its application and limitations in the investigation of thyroid disease. Med Sci Monit. 2022;28:e937738
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ. 2015;350:g7594
15. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441
16. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part II—binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96
17. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I—continuous outcomes. Stat Med. 2019;38(7):1262-75
18. Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R^2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. Stat Med. 2021;40(4):859-64
19. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74
20. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36(11):1-13
21. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019;134:93-101
22. Maurya NS, Kushwah S, Kushwaha S, et al. Prognostic model development for classification of colorectal adenocarcinoma by using machine learning model based on feature selection technique Boruta. Sci Rep. 2023;13(1):6413

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

23. O'Connell NS, Jaeger BC, Bullock GS, Speiser JL. A comparison of random forest variable selection methods for regression modeling of continuous outcomes. Brief Bioinform. 2025;26(2):bbaf096
24. Yang M, Sun M, Jiang C, et al. Thyroid hormones and carnitine in the second trimester negatively affect neonate birth weight: A prospective cohort study. Front Endocrinol (Lausanne). 2023;14:1080969
25. Vuong ADB, Tran NH, Pham TH, et al. Soluble fibrin monomer complex and D-dimer concentrations between patients at low and high risk of venous thromboembolism before delivery according to RCOG score assessment: An observational study among 100 third-trimester Vietnamese pregnancies. J Clin Med. 2025;14(5):1399
26. Nguyen Tran TN, Nguyen HT, Cao NT, et al. Umbilical cord coiling index in predicting neonatal outcomes: A single-center cross-sectional study from Vietnam. J Matern Fetal Neonatal Med. 2025;38(1):2517763
27. Liu Y, Li G, Guo N, et al. Association between maternal characteristics and the risk of isolated maternal hypothyroxinemia. Front Endocrinol (Lausanne). 2022;13:843324
28. Zhang E, Zhang Z, Chen G, et al. Associations of ambient particulate matter with maternal thyroid autoimmunity and thyroid function in early pregnancy. Environ Sci Technol. 2024;58(21):9082-90
29. Cai L, Wang P, Xue C, et al. Clinical characteristics and risk factors associated with adverse pregnancy outcomes in patients with gestational hypothyroidism: A case-control study. Endocr Pract. 2024;30(2):101-6
30. Hu M, Asami C, Iwakura H, et al. Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests. Commun Med (Lond). 2022;2(1):9
31. Lang M, Binder M, Richter J, et al. mlr3: A modern object-oriented machine learning framework in R. J Open Source Softw. 2019;4(44):01903
32. Mohr F, van Rijn JN. Fast and informative model selection using learning curve cross-validation. IEEE Trans Pattern Anal Mach Intell. 2023;45(8):9669-80
33. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: A review. JAMA Psychiatry. 2020;77(5):534-40
34. Fan ZE, Lian F, Li XR. Rethinking density ratio estimation based hyper-parameter optimization. Neural Netw. 2025;182:106917
35. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Discov. 2014;28(1):92-122
36. Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. R J. 2014;6(1):79-89
37. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: What, why, how, when and where? Clin Kidney J. 2021;14(1):49-58
38. Sun Y, Yu K, Du L, et al. Application of XGBoost in the prediction of acute postoperative pain after major noncardiac surgery in older patients. Mol Pain. 2025;21:17448069251376199
39. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44-65
40. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189-232
41. Amrita C, Mitali M, Kumari CS, Rupesh K. Can yoga help to manage the symptoms of thyroid diseases? Int J Yoga. 2025;18(1):3-12
42. Chen DW, Ospina NS, Haymart MR. Social determinants of health and disparities in thyroid care. J Clin Endocrinol Metab. 2024;109(3):e1309-13

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]