Development and validation of a prediction model using sella magnetic resonance imaging-based radiomics and clinical parameters for diagnosis of growth hormone deficiency and idiopathic short stature

Kyungchul Song

The Graduate School
Yonsei University
Department of Medicine

Development and validation of a prediction model using sella magnetic resonance imaging-based radiomics and clinical parameters for diagnosis of growth hormone deficiency and idiopathic short stature

A Doctoral Dissertation
submitted to the Department of Medicine,
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Science

Kyungchul Song

June 2024

**This certifies that the Dissertation
of Kyungchul Song is approved.**

Thesis Supervisor : Hyun Wook Chae

Thesis Committee Member Ji-Hoon Na

Thesis Committee Member Young-Jun Rhie

Thesis Committee Member Hyun Joo Shin

Thesis Committee Member Taehoon Ko

**The Graduate School
Yonsei University
June 2024**

ACKNOWLEDGEMENTS

I extend my sincere gratitude to Professor Hyun Wook Chae for his exceptional guidance and unwavering support during my pursuit of a doctoral degree in pediatric endocrinology. His expertise, mentorship, and encouragement have been instrumental in shaping my academic and research endeavors. Professor Chae's dedication to fostering a rich learning environment has not only facilitated my academic growth but has also inspired me to strive for excellence in my chosen field.

A special acknowledgment goes to Professor Ho-Seong Kim, my mentor, who imparted the foundational knowledge of pediatric endocrinology. His insightful perspectives and guidance have significantly enriched the development of my ideas and the overall quality of my work. Additionally, I express my heartfelt thanks to Dr. Beomseok Sohn for his valuable contributions, which have been crucial in making basic concept of my research.

Beyond my academic mentors, I want to acknowledge and express deep gratitude to my family, especially my mother. Her constant support, prayers, and love have been the cornerstone of my journey. I am truly fortunate to have such a strong and caring foundation that has

propelled me forward in both my academic and personal life. My family's unwavering belief in my capabilities has been a driving force, encouraging me to overcome challenges and pursue my passion in pediatric endocrinology.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development and validation of a prediction model using sella magnetic resonance imaging-based radiomics and clinical parameters for diagnosis of growth hormone deficiency and idiopathic short stature

For differential diagnosis of growth hormone deficiency (GHD) and idiopathic short stature (ISS), growth hormone provocation test is the gold standard test, but it is very invasive and has limited validity and reproducibility. Thus, investigations on prediction model for differential diagnosis of GHD and ISS are required.

Radiomics, a method of extracting various features using mathematical algorithm, can find molecular profile and disease characteristics which cannot be detected by human eye. Based on concept of information in biomedical images which reflects underlying pathophysiology, radiomics converts digital medical images into mineable high-dimensional data. However, investigations on pituitary gland using radiomics in children are limited.

Therefore, we aim to develop a machine learning-based prediction model for diagnosis of GHD and ISS using radiomics including sella magentic resonance imaging (MRI) and clinical parameters and validate the prediction model with external validation.

A total of 293 children with normal sella MRI findings in the training set and 47 children in the test set from different hospitals were enrolled. A total of 186 radiomic features were extracted from the pituitary glands for both the T2-weighted and contrast-enhanced T1-image. The clinical parameters included auxological data, insulin-like growth factor-I (IGF-I), and bone age. The XGBoost algorithm was used to train the prediction models. Internal validation was conducted using five-fold cross-validation on the training set, and external validation was conducted on the test set. Model performance was assessed by plotting the area under the receiver operating characteristic curve (AUC). The mean absolute Shapley values were computed to quantify the impact of each parameter.

The AUCs of the clinical, radiomics, and combined models were 0.684, 0.691, and 0.830, respectively, in the external validation. Among the clinical parameters, the major contributing factors to prediction were body mass index standard deviation score (SDS), chronological age–bone age, weight SDS, growth velocity, and IGF-I SDS in the clinical model. Among the radiomics features, Inverse Variance from T2 Weighted image and Energy from contrast-enhanced T1-image were the major factors contributing to the radiomics model. In the combined model, radiomic features added incremental value to the prediction.

In conclusion, this study underscores the potential of radiomics-based diagnostic models to overcome the limitations of conventional modalities for diagnosis of GHD and ISS. These findings substantiate the pivotal roles of radiomics and machine learning in pediatrics into new enhanced diagnostics.

Key words : growth hormone deficiency, idiopathic short stature, machine learning, radiomics

1. INTRODUCTION

Short stature is defined as a height less than the third percentile or more than two standard deviations below the average height for those of the same age, sex, and ethnicity. In children, short stature is related to medical problems and lifestyle including exercise and diet as well as psychosocial performance.[1-3] Short stature is usually a normal variation among the population, Negative stereotypes related to short stature is prevalent although children with chief complaint of short stature is within normal range, which may lead to psychosocial problems.[3,4] Among etiologies of short stature, growth hormone (GH) deficiency (GHD) and idiopathic short stature (ISS) account for the most common causes.[2] GH, a peptide hormone secreted from pituitary gland, stimulates linear bone growth and cell reproduction, and GHD is defined as a condition induced by insufficient secretion of GH.[1,4] Whereas, ISS is defined as short stature without evidence of systemic, endocrine, nutritional, or chromosomal abnormalities.[1,4]

For the diagnosis of GHD, meticulous evaluation, including the measurement of anthropometric data, bone age, insulin-like growth factor I (IGF-I), and GH provocation tests, is required, among which the GH provocation test is considered the gold standard.[5,6] GHD can be diagnosed in children with short stature who show insufficient GH levels after at least two GH provocation tests. However, the GH provocation test is extremely invasive and burdensome to patients and requires hospitalization and multiple blood samplings; therefore, investigations on noninvasive screening methods to replace GH provocation test are required.[7]

Etiologies of GHD include pathological causes such as brain tumor and hypoxic brain damage; therefore, sella magnetic resonance imaging (MRI) is required for the evaluation of GHD.[2] Several studies investigated difference of pituitary volume in sella MRI according to etiologies of short stature, and Kessler et al. reported that pituitary volume is different between children with GHD and ISS and increases with increase in age.[8-11]

Meanwhile, artificial intelligence (AI) is increasingly being utilized as a novel approach in research and diagnosis involving medical imaging. In the landscape of medical imaging, the journey of AI began in 1955 with John McCarthy's definition that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.[12] This vision gave rise to supervised machine learning, the cornerstone of the radiological AI, where algorithms are trained to recognize pathologies such as tumors in computed tomography (CT) or

MRI scans based on gold standard markers.[12,13] These algorithms refine their diagnostic prowess by learning from numerous cases and then applying this knowledge to identify these markers in new test cohort with unseen images.

However, classical AI in imaging analysis has limitations.[14] Information in the diagnostic process remains hidden within the computational black box as well as just a simple answer such as presence of a lesion is provided. Thus, radiomics a method of extracting various features using mathematical algorithms, emerges as a revolutionary technique addressing these shortcomings by offering a quantitative image analysis framework.[13,15] Radiomics can be used to determine molecular profiles and disease characteristics that cannot be detected by the human eye.[14,16] Based on the concept of information in biomedical images that reflects the underlying pathophysiology, radiomics converts digital medical images into mineable high-dimensional data.[14] New imaging biomarkers generated by radiomics can be integrated with clinical data such as genetic information, and overcome limitation of predicting diseases with clinical parameters.[15]

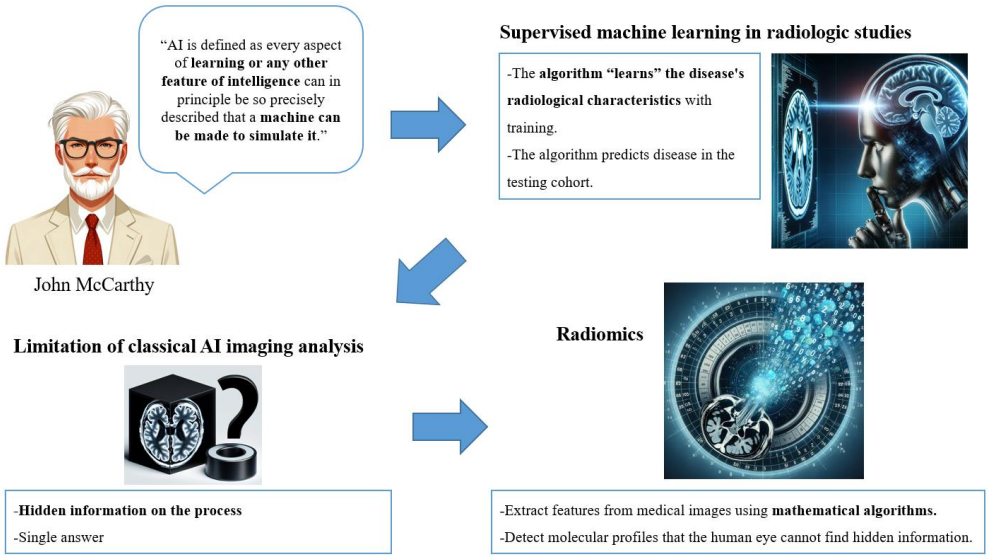

Figure 1. Development AI in radiology and radiomics

Abbreviation: AI, artificial intelligence.

AI imaging analysis has been investigated also in pediatrics.[13] In a systematic review, 5 of 6 studies showed that AI was superior to human experts for diagnosis of tumor.[17] Our previous study demonstrated the program developed using AI showed a high level of precision in determining bone age and final adult height for Korean youths.[18] However, investigations and applications on AI imaging analysis in children and adolescents are still limited compared to that in adult. Moreover, although quantitative analyses of medical images have been performed in adults because numerous radiomic features can be extracted and analyzed using radiomics, investigations of the pituitary gland using radiomics in children are limited.[14,19,20]

Notably, clinical parameters associated with GHD diagnosis have been investigated in several studies that included anthropometric data, such as height and body mass index (BMI), and laboratory tests such as IGF-I.[21-24] However, literature regarding prediction models for the differential diagnosis of GHD and ISS in children are currently lacking, especially regarding prediction models using both radiomics and clinical parameters.

Therefore, this study aimed to develop a machine-learning-based prediction model for the diagnosis of GHD and ISS using radiomics and clinical parameters. Our objectives were to (1) extract radiomic features using T2-weighted image (T2WI) and contrast enhanced T1-weighted image (T1C) in sella MRI; (2) develop a prediction model using radiomic features and clinical parameters; (3) compare predictability among the models using radiomics, clinical parameters, and both parameters; (4) estimate the accuracy of the predictive models with external validation; and (5) evaluate the contribution of each clinical parameter and radiomic feature from the prediction models. To achieve this goal, we investigated the following contents: (1) baseline characteristics of the participants; (2) ROC curve analyses of clinical, radiomics, and combined models; and (3) Shapley value of clinical parameters and radiomic features

2. MATERIALS AND METHODS

2.1.Study population (Figure 2)

To develop the prediction model for diagnosis of GHD and ISS, electronic records of children aged 18 years or younger with short stature who underwent GH provocation test and sella MRI between March 2011 and July 2020 were extracted from the Severance Clinical Research Analysis Portal of Severance Hospital. Among these participants, subjects with endocrinological or systemic pathology

such as skeletal dysplasia, small for gestational age, genetic disease including chromosomal abnormality, Russel-Silver syndrome, and Prader-Willi syndrome, and chronic disease including cancer, congenital heart disease, and systemic lupus erythematosus and/or those with pituitary lesion including tumor and empty sella were excluded for the final derivation set. For external validation set, electronic records of children aged 18 years or younger with short stature who underwent GH provocation test and sella MRI between September 2020 and November 2022 were extracted from the Severance Clinical Research Analysis Portal of Yongin Severance Hospital. Exclusion criteria for final external validation set was same with those for derivation set. Finally, a total of 293 children and adolescents with sella MRI in the training set and 47 children and adolescents in the test set were enrolled.

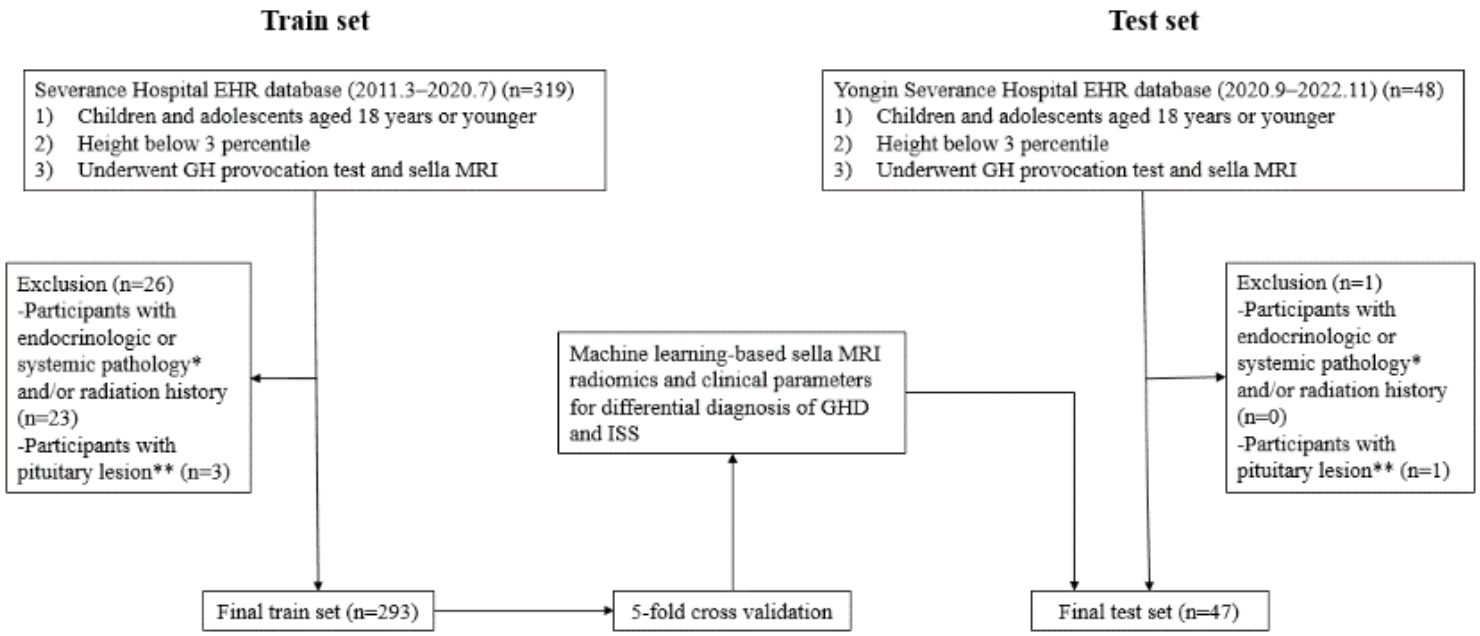
Train set
Test set
Severance Hospital EHR database (2011.3–2020.7) (n=319)
1) Children and adolescents aged 18 years or younger
2) Height below 3 percentile
3) Underwent GH provocation test and sella MRI
Yongin Severance Hospital EHR database (2020.9–2022.11) (n=48)
1) Children and adolescents aged 18 years or younger
2) Height below 3 percentile
3) Underwent GH provocation test and sella MRI
Exclusion (n=26)
-Participants with endocrinologic or systemic pathology* and/or radiation history (n=23)
-Participants with pituitary lesion** (n=3)
Exclusion (n=1)
-Participants with endocrinologic or systemic pathology* and/or radiation history (n=0)
-Participants with pituitary lesion** (n=1)
Machine learning-based sella MRI radiomics and clinical parameters for differential diagnosis of GHD and ISS
Final train set (n=293)
5-fold cross validation
Final test set (n=47)

Figure 2. Study flow.

* Endocrinological or systemic pathology: hypopituitarism, adrenal insufficiency, hyperthyroidism or hypothyroidism except euthyroid state, skeletal dysplasia, small for gestational age, genetic disease including chromosomal abnormalities, Russel-Silver syndrome, Prader-Willi syndrome, and chronic diseases including cancer, including a history of brain irradiation, congenital heart disease, and systemic lupus erythematosus.

**Pituitary lesion: pituitary tumor and/or empty sella

Abbreviation: EHR; electronic health record, GH; growth hormone, MRI; magnetic resonance imaging, GHD; growth hormone deficiency, ISS; idiopathic short stature.

2.2. Definition of GHD and ISS

GHD was defined as follows: (1) height below the third percentile for age, sex, and race based on the 2017 Korean National Growth Charts;[25] (2) peak GH level below 10 ng/mL after stimulation in two types of GH provocation tests using insulin, arginine, and/or L-dopa; and (3) children without genetic, endocrine, or systemic abnormalities.[2,26]

Among the participants with GHD, severe GHD was defined as follows: (1) height below the third percentile for age, sex, and race based on the 2017 Korean National Growth Charts;[25] (2) peak GH level below 5 ng/mL after stimulation in two types of GH provocation tests using insulin, arginine, and/or L-dopa; and (3) children without genetic, endocrine, or systemic abnormalities.[27]

ISS was defined as height below the third percentile for individuals of the same age, sex, and race, with no other identifiable causes, including genetic, endocrine, or systemic pathologies.[2,25,26]

2.3. Clinical parameters

Height was recorded with an accuracy of 0.1 cm, whereas body weight was measured using an electronic load with a precision of 0.01 kg. BMI was calculated by dividing body weight in kilograms by the square of height in meters (kg/m^2). Height, weight, and BMI were expressed as standard deviation scores (SDS) using the 2017 Korean National Growth Charts.[25] Children were categorized based on their BMI into three groups: normal (<85th percentile), overweight (85th–95th percentile), or obese (≥95th percentile). Mid-parental height (MPH) was determined by calculating the average height of the parents and adjusting it by subtracting 6.5 cm for girls and adding 6.5 cm for boys. Puberty was considered at any pubertal development with Tanner stage ≥ 2.[28,29]

Growth hormone (GH) and insulin-like growth factor I (IGF- I) levels were determined using a chemiluminescence immunoassay using a LIAISON® XL immunoassay system (DiaSorin, S.p.A., Saluggia, Italy) using the human GH (hGH) reagent traceable to World Health Organization (WHO) 2nd International Standard 97/574 and the IGF-I reagent traceable to WHO 1st International Standard IGF-1 National Institute for Biological Standards and Control (NIBSC) code 02/254, respectively, in Severance Hospital. IGF binding protein 3 (IGFBP-3) was determined by an immunoradiometric assay using the IGFBP-3 immunoradiometric assay reagent (Immunodiagnostic Systems, UK).

In Yongin Severance Hospital, GH, IGF-I, and IGFBP-3 levels were determined using a electrochemiluminescence immunoassay on cobas® e801 immunoassay system (Roche Diagnostics

GmbH, Mannheim, Germany). GH was determined using Elecsys hGH reagent traceable to the international reference preparations, NIBSC code 98/574, at Seoul Clinical Laboratories, as a send-out test for the patients. Serum levels of IGF-I and IGFBP-3 were determined using the Elecsys IGF-1 reagent standardized against WHO 02/254 internal standards and Elecsys IGFBP-3 reagent standardized against IDS iSYS® IGFBP-3, respectively.

SDS values of IGF-I and IGF binding protein 3 (IGFBP-3) were calculated based on reference data for the Korean population.[30] Bone age was assessed according to the Greulich-Pyle method by experienced pediatric endocrinologists.[31] In addition, we calculated chronological age–bone age (CA–BA).

2.4. Image acquisition

A. The training set:

The participants were scanned using various 3.0 T magnetic resonance imaging (MRI) units (Achieva; Philips Medical Systems, Amsterdam, the Netherlands). Coronal view of T2-weighted image (T2WI) and contrast-enhanced T1-weighted image (T1C) which are the most informative representative series in sella MRI were included in the imaging protocols. The sequence parameters of the T2WI and T1C were as following: TR/TE=2129/90 ms; slice thickness=1.0 mm; intersection gap=0 mm; field of view (FOV)=36×24 cm; flip angle = 90; pixel spacing=0.188×0.188 mm; and TR/TE = 2000/10 ms; slice thickness=1.2 mm; intersection gap = 0 mm; FOV= 32 × 24 cm; flip angle = 90; and pixel spacing = 0.391 × 0.391 mm.

B. The test set:

MRI data were acquired using a 3T MRI scanner (Ingenia Elition X or Ingenia CX, Philips Healthcare, Best, the Netherlands) with a 32-channel head coil. The imaging protocols included coronal view of T2-weighted image (T2WI) and contrast-enhanced T1-weighted image (T1C) which is representative series in sella MRI. The sequence parameters of the T2WI and T1C are as following: TR/TE=2132/80 ms; slice thickness=1.5 mm; spacing between slices=1.5 mm; FOV=20×20 cm; flip angle = 90; pixel spacing=0.391×0.391 mm and TR/TE = 608/12 ms; slice thickness=1.5 mm; spacing between slices=1.5 mm; FOV= 20 x 20 cm; flip angle = 90; pixel spacing = 0.391 x 0.391 mm, respectively.

2.5. Image processing and radiomics feature extraction

The T2WI and T1C from the sella MRI were examined, and the entire pituitary gland was identified within the region of interest. The outermost borderline of the sliced pituitary gland was outlined.

Following the conversion of the T2WI and T1C from the sella MRI, which were in Digital Imaging and Communication in Medicine format, into NIfTI files, the images were resampled to a resolution of 1×1×1 mm. In addition, N4 bias correction was used for a correction of low-frequency intensity non-uniformity.[32] The images were performed by a radiologist with a 10-year experience, who was unaware of the participants' clinical information. An open-source software (Medical Image Processing, Analysis, and Visualization; Center for Information Technology, National Institutes of Health, Bethesda, MD, USA) was used for the analysis. Region of the interest of the pituitary gland in each image slice was obtained semi-automatically using edge detection, signal intensity thresholding, and region growing. To ensure the reliability of the segmentation, another radiologist with 10 years of experience independently conducted the segmentation of 10% of the final images selected from the dataset, which were chosen randomly. The Dice coefficient was calculated to assess the agreement between the segmentation masks generated by the two radiologists. Next, Pyradiomics 2.1.0 (http://www.radiomics.io/pyradiomics.html) was used for extraction of the radiomic feature with 128 fixed bin counts.[33] Finally, we extracted 14 shape, 18 first-order, 24 gray-level co-occurrence matrix (GLCM), 16 gray-level run length matrix (GLRLM), 16 gray-level size zone matrix (GLSZM), and 5 neighborhood gray tone difference matrix (NGTDM) from the region of interests on pituitary gland. In total, 186 radiomic features were extracted from T2WI and T1C, respectively.

2.6. Machine learning and statistical analysis (Figure 3)

We trained and compared three models that classified GHD and ISS according to the following parameters: radiomic features, clinical parameters, and both of these parameters. This comprehensive approach aimed to assess the combined predictive ability of radiomics and clinical parameters for diagnosis.[34,35] For training of the models, The XGBoost algorithm was performed. XGBoost is an ensemble of decision trees with high predictive and explanatory ability.[36] In particular, XGBoost can learn datasets with missing values. Bayesian optimization was used for optimizing the XGBoost. We performed internal validation using repeated fivefold cross-validation. The evaluation metrics used were accuracy, sensitivity, specificity, precision, and area under the receiver operating

characteristic (ROC) curve (AUC). The bootstrap method was used for the pairwise comparison of the AUC, and the prediction models were externally validated using the Yongin Severance Hospital dataset. All analyses were performed using Python 3.9. Statistically significance was defined as *p* value less than 0.05.

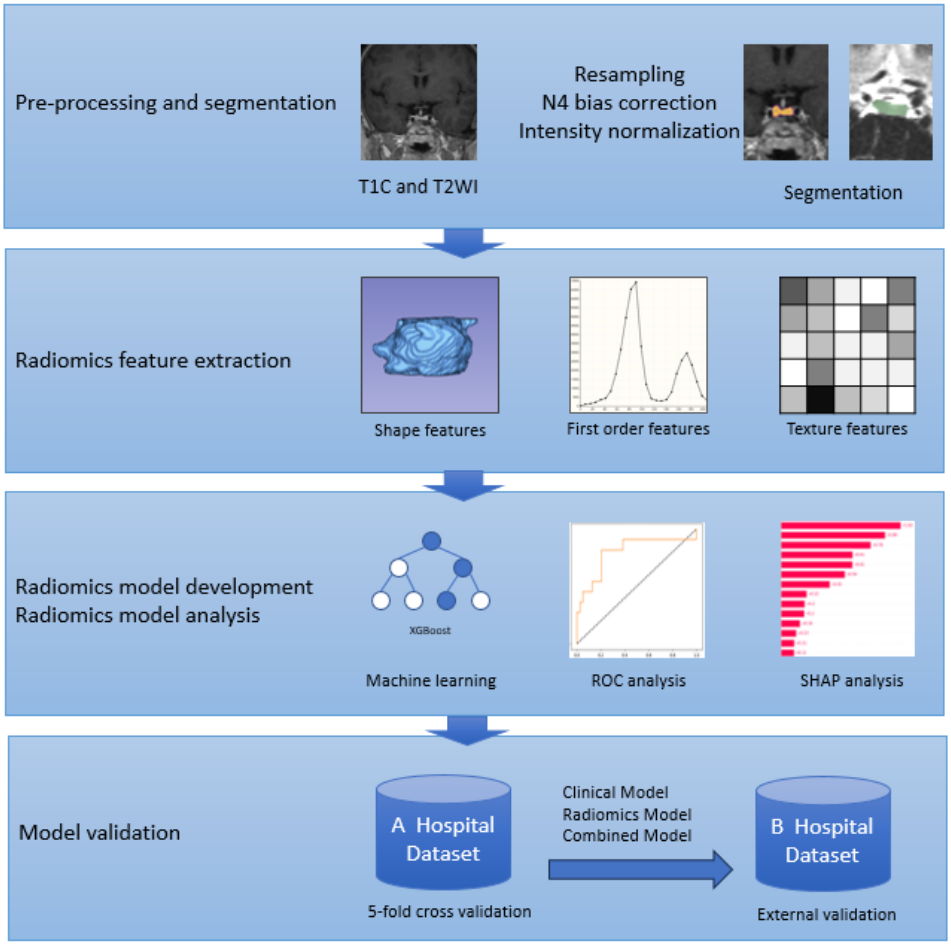

Figure 3. Machine learning pipeline.

Abbreviation: T1C; contrast-enhanced T1-weighted image, T2WI; T2-weighted image, ROC; receiver operating characteristics, SHAP; Shapley additive explanations.

2.7. Interpretation of the model using Shapley Additive exPlanations (SHAP)

SHAP was used for evaluation and interpretation the significance of the clinical parameter and radiomic parameters from the prediction models.[37] Contribution of the features was measured with SHAP, called the Shapley value, to the prediction of GHD. This analysis allowed us to visualize and understand the significance of each feature in contributing to the performance of the model. This study used three perspectives to interpret the models: feature importance plots, dot summary plots, and waterfall plots. Importance was calculated by averaging the Shapley values per feature. The dot summary plot is a scatter plot of the feature importance based on the magnitude of each feature value. The waterfall plot shows the impact of the features on the machine learning models for each case. This study sampled true-positive and true-negative cases for GHD classification and examined a machine learning model using waterfall plots.

3. RESULTS

3.1. Baseline characteristics of the subjects according to the etiology of short stature (Table 1)

BMI and the proportions of underweight, overweight, and obese participants were higher in participants with GHD than in those with ISS. IGF-I and IGF-I SDS were lower in participants with GHD than in those with ISS, whereas CA–BA was higher in those with GHD.

Table 1. Baseline characteristics of the subjects according to etiology of short stature

	GHD (n = 248)	ISS (n = 96)	p-value
Sex (male)	150 (60.5)	53 (55.2)	0.365
Age, year	7.24 ± 2.81	7.21 ± 2.73	0.917
Height, cm	112.53 ± 14.88	111.68 ± 14.34	0.630
Height SDS	-2.54 ± 0.56	-2.67 ± 0.67	0.075
Weight, kg	20.94 ± 7.73	19.37 ± 6.17	0.071
Weight SDS	-2.10 ± 2.31	-2.42 ± 1.02	0.197

BMI, kg/m^2	16.06 ± 2.50	15.16 ± 1.72	0.001
BMI SDS	-0.73 ± 2.53	-1.05 ± 1.00	0.221
BMI percentile			0.023
underweight	189 (76.2)	69 (71.9)	
normal	40 (16.1)	24 (25.0)	
overweight	11 (4.4)	2 (2.1)	
obesity	8 (3.2)	1 (1.0)	
Growth velocity, cm/year	4.44 ± 1.63	4.21 ± 1.56	0.252
Pubertal status			0.294
prepuberty	213 (85.9)	78 (81.2)	
puberty	35 (14.1)	18 (18.8)	
MPH SDS	-0.09 ± 0.08	-0.09 ± 0.09	0.510
MPH SDS - Height SDS	2.63 ± 0.57	2.76 ± 0.67	0.080
IGF-I, ng/mL	137.55 ± 58.38	153.58 ± 70.16	0.030
IGF-I SDS	-0.79 ± 0.63	-0.69 ± 0.71	0.015
IGFBP-3, ng/mL	2344.02 ± 1127.25	2159.60 ± 786.76	0.317
IGFBP-3 SDS	0.82 ± 0.83	0.68 ± 0.76	0.001
Bone age, year	6.69 ± 2.76	6.72 ± 2.72	0.941
CA–BA, year	0.61 ± 0.95	0.34 ± 0.97	0.031

Categorical variables are shown as numbers (percentages) and continuous variables are shown as mean ± standard deviation. P-values for categorical variables are determined using chi-square tests and for continuous data are determined using independent t-tests.

Abbreviation: GHD; growth hormone deficiency, ISS; idiopathic short stature, SDS; standard deviation score, BMI; body mass index, MPH; mid-parental height, IGF-I; insulin-like growth factor I, IGFBP-3; insulin-like growth factor binding protein-3, CA–BA; chronological age–bone age.

3.2. Baseline characteristics of the training set and the test set (Table 2)

Regarding the baseline characteristics of the participants in the training and test sets, the

proportions of boys, underweight, prepuberty, and ISS were higher in the training set than in the test set. Age, height, MPH SDS, and BA were higher in the test set than in the training set, whereas the MPH SDS height, SDS, and CA–BA were higher in the training set. In the training set, the proportion of GHD, ISS, and severe GHD were 69.6%, 30.4%, and 34.3%, respectively. In the test set, the corresponding values were 87.5%, 12.5%, and 45.8%, respectively.

Table 2. Baseline characteristics of the training set and the test set

	Training set (n = 296)	Test set (n = 48)	p-value
Sex (male)	182 (61.5)	27 (56.2)	0.020
Age, year	7.04 ± 2.81	8.44 ± 2.31	0.001
Height, cm	110.55 ± 15.50	121.18 ± 13.55	<0.001
Height SDS	-2.77 ± 2.18	-2.15 ± 0.53	0.053
Weight, kg	-2.77 ± 2.18	-2.15 ± 0.53	0.053
Weight SDS	-2.26 ± 2.23	-1.56 ± 0.75	0.032
BMI, kg/m^2	15.68 ± 2.30	16.57 ± 2.46	0.080
BMI SDS	-0.87 ± 2.34	-0.49 ± 1.05	0.110
BMI percentile			0.033
underweight	60 (20.3)	4 (8.3)	
normal	217 (73.3)	39 (81.2)	
overweight	11 (3.7)	2 (4.2)	
obesity	8 (2.7)	3 (6.2)	
Growth velocity, cm/year	4.37 ± 1.57	4.40 ± 1.87	0.100
Pubertal status			0.046
prepuberty	255 (86.1)	36 (75.0)	
puberty	41 (13.9)	12 (25.0)	
MPH SDS	-0.04 ± 0.09	-0.03 ± 0.10	0.040
MPH SDS - Height SDS	2.74 ± 0.58	2.21 ± 0.54	<0.001
IGF-I, ng/mL	140.27 ± 62.56	153.32 ± 59.26	0.186
IGF-I SDS	-0.74 ± 0.69	-0.87 ± 0.36	0.207
IGFBP-3, ng/mL	1987.94 ± 718.58	4171.04 ±	<0.001

			741.61	
IGFBP-3 SDS		2.49 ± 1.15	0.51 ± 0.09	<0.001
Bone age, year		6.43 ± 2.67	8.23 ± 2.69	<0.001
CA–BA, year		0.58 ± 0.95	0.26 ± 0.98	0.040
Diagnosis				0.010
	GHD	206 (69.6)	42 (87.5)	
	ISS	90 (30.4)	6 (12.5)	
Severe GHD		72 (24.3)	22 (45.8)	<0.001

Categorical variables are shown as numbers (percentages) and continuous variables are shown as mean ± standard deviation. P-values for categorical variables are determined using chi-square tests and for continuous data are determined using independent t-tests.

Abbreviation: SDS; standard deviation score, BMI; body mass index, MPH; mid-parental height, IGF-I; insulin-like growth factor I, IGFBP-3; insulin-like growth factor binding protein-3, CA–BA; chronological age–bone age, GHD; growth hormone deficiency, ISS; idiopathic short stature.

3.3. ROC curve analyses of clinical, radiomics, and combined models (Table 3, 4, and 5 and Figure 4 and 5)

Table 3 and Figure 4 and 5 summarize the results of the ROC curve analyses and present the AUCs with corresponding 95% confidence intervals (CIs) for GHD prediction using the clinical, radiomics, and combined models. Among the clinical parameters, age, sex, height SDS, weight SDS, BMI SDS, growth velocity, pubertal state, MPH SDS, MPH SDS – height SDS, IGF-I SDS, and CA–BA were assessed using clinical and combined models. IGFBP-3 was excluded from the parameters because the value was substantially different between the two centers owing to different assays and reagents.

The accuracy and AUC (95% CI) of the clinical model were 0.717 and 0.690 (0.628–0.753) and 0.702 and 0.684 (0.590–0.778) for internal and external validation, respectively. In the radiomics model, the corresponding values were 0.668 and 0.674 (0.609–0.738) for internal validation and 0.698 and 0.691 (0.620–0.762) for external validation. In the combined model, the corresponding values were 0.817 and 0.835 (0.776–0.896) for internal validation and 0.813 and 0.830 (0.741–0.919) for external validation.

Table 3. AUCs of each model for predicting GHD

		Accuracy	Sensitivity	Specificity	Precision	AUC (95% CI)
Clinical model	Internal validation	0.717	0.738	0.667	0.838	0.690 (0.628–0.753)
	External validation	0.702	0.707	0.667	0.936	0.684 (0.590–0.778)
Radiomics model	Internal validation	0.678	0.691	0.685	0.578	0.674 (0.609–0.738)
	External validation	0.698	0.643	0.667	0.831	0.691 (0.620–0.762)
Combined model	Internal validation	0.817	0.857	0.722	0.878	0.835 (0.776–0.896)
	External validation	0.813	0.810	0.833	0.971	0.830 (0.741–0.919)

P-values are determined using ROC curve for AUC.

Abbreviation: AUC; area under the receiver operating characteristics curve, GHD; growth hormone deficiency, CI; confidence interval, ROC; receiver operating characteristics.

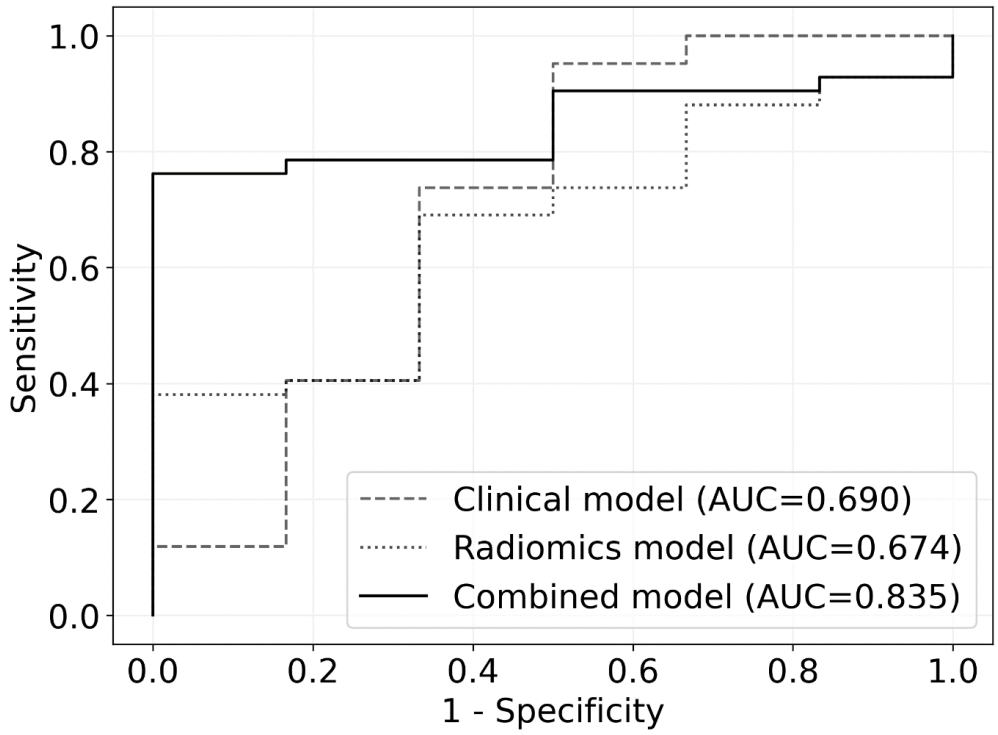

Figure 4. ROC curves from the clinical model, radiomics model, and combined model for internal validation.

Abbreviation: AUC; area under the receiver operating characteristics curve, ROC; receiver operating characteristics.

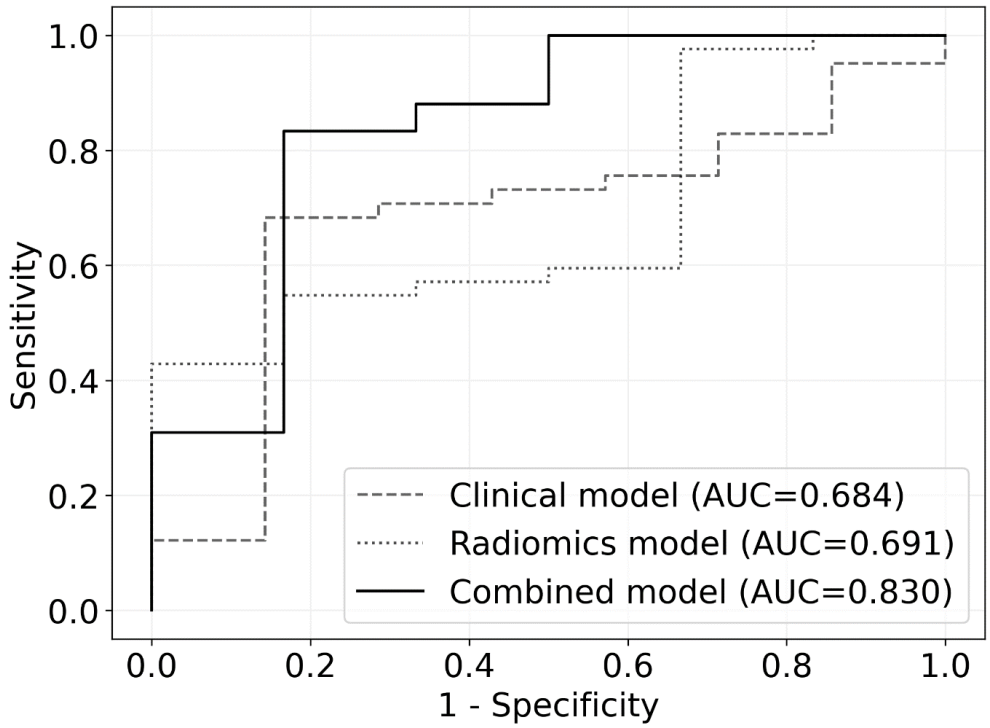

Figure 5. ROC curves from the clinical model, radiomics model, and combined model for external validation.

Abbreviation: AUC; area under the receiver operating characteristics curve, ROC; receiver operating characteristics.

In the pairwise comparison among the prediction models for GHD, the combined model outperformed both the clinical and radiomics models in internal validation (combined model vs. clinical model, $p = 0.012$; combined model vs. radiomic model, $p = 0.026$) and external validation (combined model vs. clinical model, $p = 0.034$; combined model vs. radiomic model, $p = 0.019$) (Table 4). The AUC were not statistically different between the clinical and radiomic models.

Table 4. Comparison of AUCs of prediction models for GHD

	Clinical model	Radiomics model	Combined model
Internal validation			
Clinical model	Reference	0.208	0.012
Radiomics model	0.208	Reference	0.026
Combined model	0.012	0.026	Reference
External validation			
Clinical model	Reference	0.284	0.034
Radiomics model	0.284	Reference	0.019
Combined model	0.034	0.019	Reference

The bootstrap method was used to perform pairwise comparisons between AUCs for the variables.
Abbreviation: AUC; area under the receiver operating characteristics curve, GHD; growth hormone deficiency.

Table 5 summarizes the results of the ROC curve analyses and present the AUCs with corresponding 95% confidence intervals (CIs) for severe GHD prediction using the clinical, radiomics, and combined models.

The accuracy and AUC (95% CI) of the clinical model were 0.629 and 0.650 (0.592–0.708) and 0.634 and 0.651 (0.597–0.705) for internal and external validation, respectively. In the radiomics model, the corresponding values were 0.640 and 0.655 (0.598–0.713) for internal validation and 0.630 and 0.648 (0.598–0.698) for external validation. In the combined model, the corresponding values were 0.651 and 0.665 (0.604–0.725) for internal validation and 0.645 and 0.672 (0.613–0.731) for external validation.

Table 5. AUCs of each model for predicting severe GHD

		Accuracy	Sensitivity	Specificity	Precision	AUC (95% CI)
Clinical model	Internal validation	0.629	0.642	0.591	0.827	0.650 (0.592–0.708)
	External validation	0.634	0.625	0.678	0.688	0.651 (0.597–0.705)
Radiomics model	Internal validation	0.640	0.657	0.591	0.830	0.655 (0.598–0.713)
	External validation	0.630	0.615	0.670	0.667	0.648 (0.598–0.698)
Combined model	Internal validation	0.651	0.672	0.591	0.833	0.665 (0.604–0.725)
	External validation	0.645	0.621	0.701	0.700	0.672 (0.613–0.731)

P-value determined using ROC curve for AUC.

Abbreviation: AUC; area under the receiver operating characteristics curve, GHD; growth hormone deficiency, CI; confidence interval, ROC; receiver operating characteristics.

In the pairwise comparison among the prediction models for severe GHD, the combined model outperformed the clinical model in internal validation (combined model vs. clinical model, $p = 0.045$) (Table 6). In external validation, the combined model outperformed both the clinical and radiomics

models in internal validation (combined model vs. clinical model, $p = 0.031$; combined model vs. radiomic model, $p = 0.048$). The AUC were not statistically different between the clinical and radiomic models.

Table 6. Comparison of AUCs of prediction models for severe GHD

	Clinical model	Radiomics model	Combined model
Internal validation			
Clinical model	Reference	0.201	0.045
Radiomics model	0.201	Reference	0.076
Combined model	0.045	0.076	Reference
External validation			
Clinical model	Reference	0.312	0.031
Radiomics model	0.312	Reference	0.048
Combined model	0.031	0.048	Reference

The bootstrap method was used to perform pairwise comparisons between AUCs for the variables.
Abbreviation: AUC; area under the receiver operating characteristics curve, GHD: growth hormone deficiency.

3.4. Shapley value of clinical parameters and radiomic features (Figure 6−13)

We computed the mean absolute Shapley values for all clinical variables and radiomic features to illustrate their contribution in the predictive models for external validation. Among the clinical parameters, the SHAP value of BMI SDS was the highest, followed by those of CA–BA, weight SDS, growth velocity, IGF-I SDS, MPH SDS, and height SDS (Figure 6). Among the radiomics features, the SHAP value of Inverse Variance from T2WI (GLCM) was the highest, followed by Energy from T1C (first order) and Sum Entropy from T2WI (GLCM) (Figure 7). In the combined model, the SHAP value of CA-BA was the highest, followed by weighted SDS, Maximum Probability from T2WI (GLCM), and Run Length Nonuniformity Normalized from T2WI (GLRLM) (Figure 8).

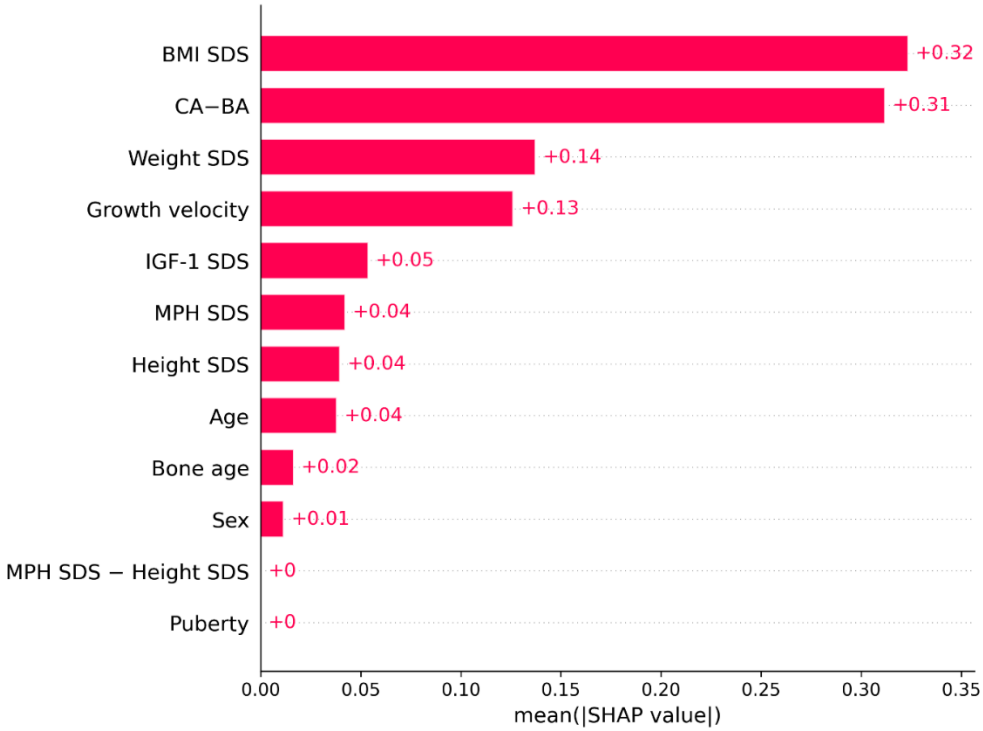

Figure 6. Mean absolute SHAP values for feature importance in the clinical model for diagnosis of GHD from external validation.

Abbreviation: BMI; body mass index, SDS; standard deviation score, CA–BA; chronological age–bone age, IGF-I; insulin-like growth factor I, MPH; mid-parental height, SHAP; Shapley additive explanations, GHD; growth hormone deficiency.

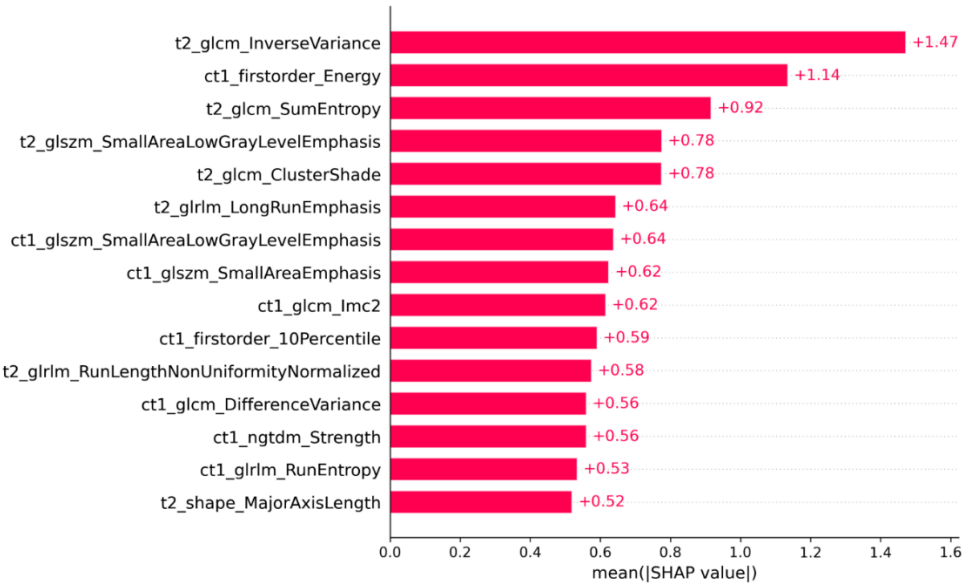

Figure 7. Mean absolute SHAP values for feature importance in the radiomics model for diagnosis of GHD from external validation.

Abbreviation: GLCM; gray-level co-occurrence matrix, GLSZM; gray-level size zone matrix, GLRLM; gray-level run length matrix, NGTDM; neighborhood gray tone difference matrix, SHAP; Shapley additive explanations, GHD; growth hormone deficiency.

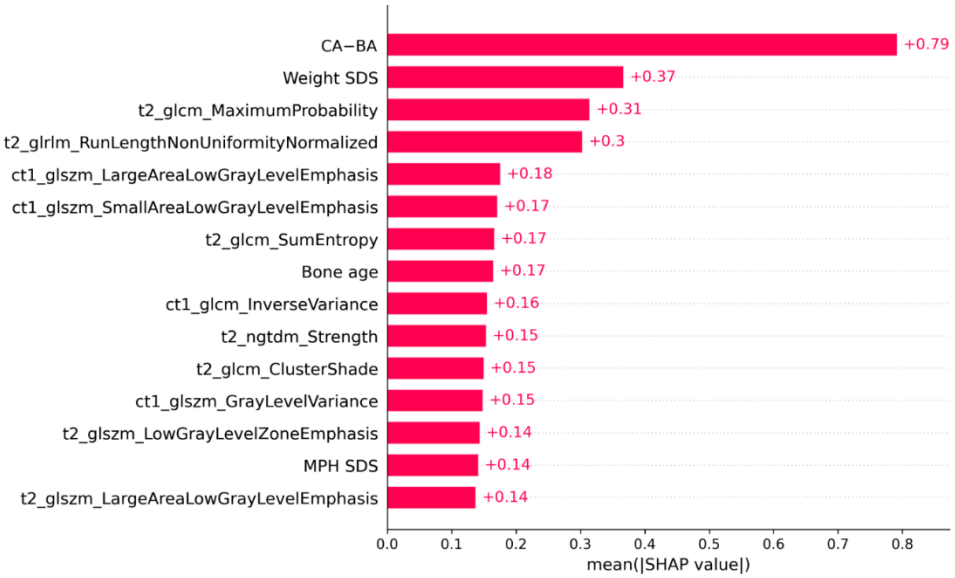

Figure 8. Mean absolute SHAP values for feature importance in the combined model for diagnosis of GHD from external validation.

Abbreviation: CA–BA; chronological age–bone age, SDS; standard deviation score, GLCM; gray-level co-occurrence matrix, GLRLM; gray-level run length matrix, GLSZM; gray-level size zone matrix, NGTDM; neighborhood gray tone difference matrix, MPH; mid-parental height, SHAP; Shapley additive explanations, GHD; growth hormone deficiency.

Analysis of the dot summary plots revealed that high CA–BA values and low value of IGF-I SDS values influenced the prediction of GHD in the clinical model (Figure 9). In the radiomics model, high value of Inverse Variance from T2WI (GLCM) influenced the prediction of the ISS, and low values of Sum Entropy from T2WI (GLCM) and Small Area Low Gray Level Emphasis from T2WI (GLSZM) influenced the prediction of GHD (Figure10). In the combined model, low values of the CA-BA influenced the prediction of GHD, whereas weight SDS, Maximum Probability from T2WI (GLCM), and Run Length NonUniformity Normalized from T2WI (GLRLM) contributed highly to the model (Figure 11).

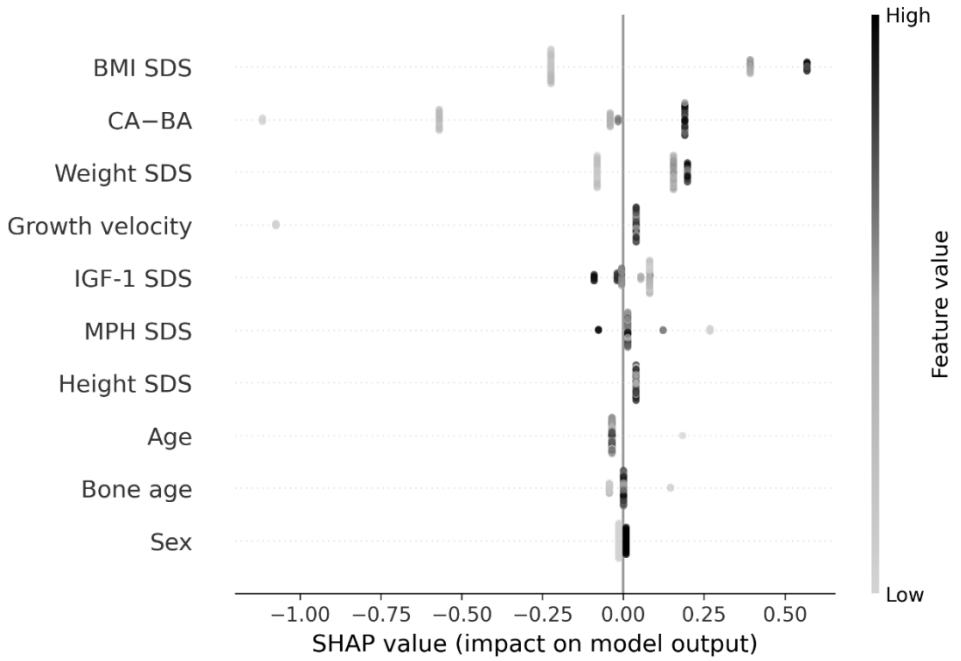

Figure 9. Dot summary plot with mean absolute SHAP value for contribution of the variables in the clinical model in external validation.

The darkness on the plot represent whether a parameter had a high or low value within the patient dataset. The horizontal position on the plot indicates whether that value had a greater or lesser impact on the prediction.

Abbreviation: BMI; body mass index, SDS; standard deviation score, CA–BA; chronological age–bone age, IGF-I; insulin-like growth factor I, MPH; mid-parental height, SHAP; Shapley additive explanations.

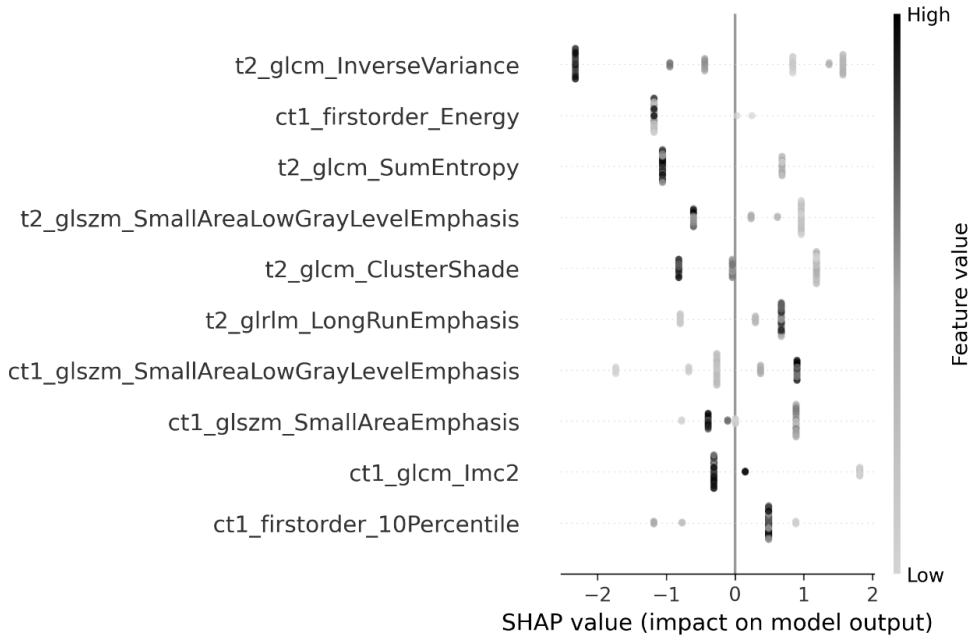

Figure 10. Dot summary plot with mean absolute SHAP value for contribution of the variables in the radiomics model in external validation.

The darkness on the plot represent whether a parameter had a high or low value within the patient dataset. The horizontal position on the plot indicates whether that value had a greater or lesser impact on the prediction.

Abbreviation: GLCM; gray-level co-occurrence matrix, GLSZM; gray-level size zone matrix, GLRLM; gray-level run length matrix, SHAP; Shapley additive explanations.

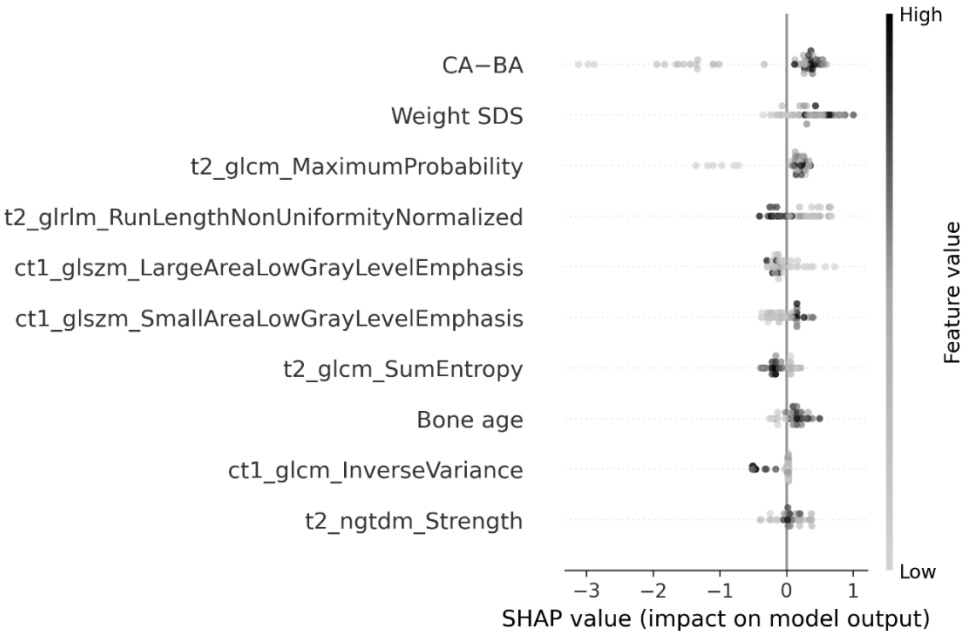

Figure 11. Dot summary plot with mean absolute SHAP value for contribution of the variables in the combined model in external validation.

The darkness on the plot represent whether a parameter had a high or low value within the patient dataset. The horizontal position on the plot indicates whether that value had a greater or lesser impact on the prediction.

Abbreviation: CA–BA; chronological age–bone age, SDS; standard deviation score, GLCM; gray-level co-occurrence matrix, GLRLM; gray-level run length matrix, GLSZM; gray-level size zone matrix, NGTDM; neighborhood gray tone difference matrix, SHAP; Shapley additive explanations.

By conducting SHAP analysis, waterfall plots were generated for each patient, and an example of such a waterfall plot using the clinical model is shown in Figure 12. The clinical model predicted the participant with ISS as ISS. In this case, the contribution of the CA–BA was the highest, followed by the BMI SDS and IGF-I SDS. Figure 13 shows a waterfall plot in which the combined model predicts a participant with GHD as having GHD. In this case, contribution of Joint Entropy from T2WI (GLCM) was the highest and those of interquartile Range from T1C (first order) was followed.

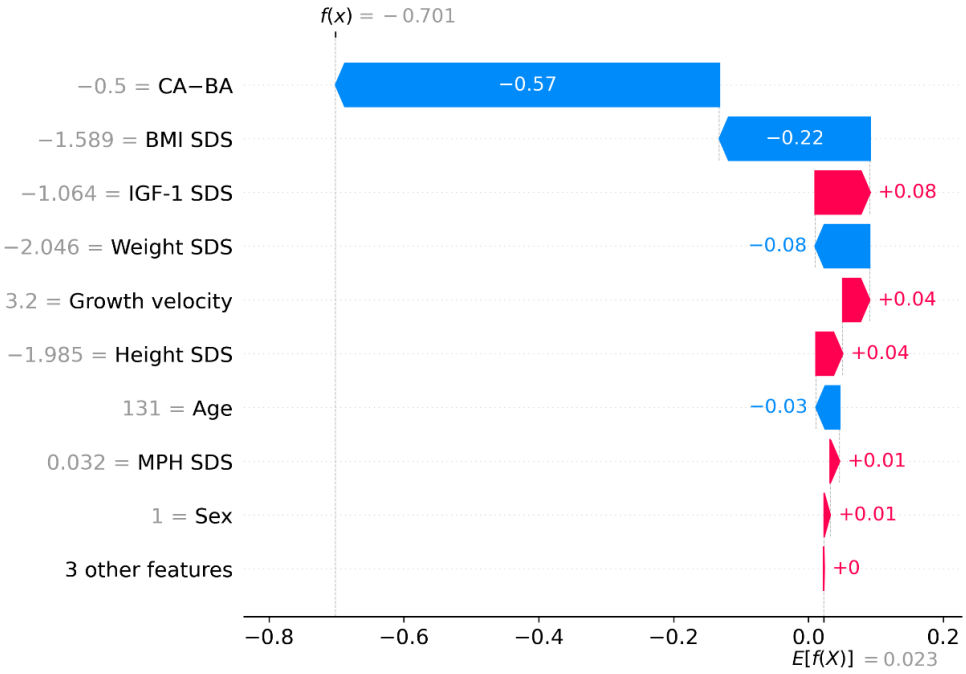

Figure 12. Representative waterfall plot of the clinical model case.

Abbreviation: CA–BA; chronological age–bone age, BMI; body mass index, SDS; standard deviation score, IGF-I; insulin-like growth factor I, MPH; mid-parental height.

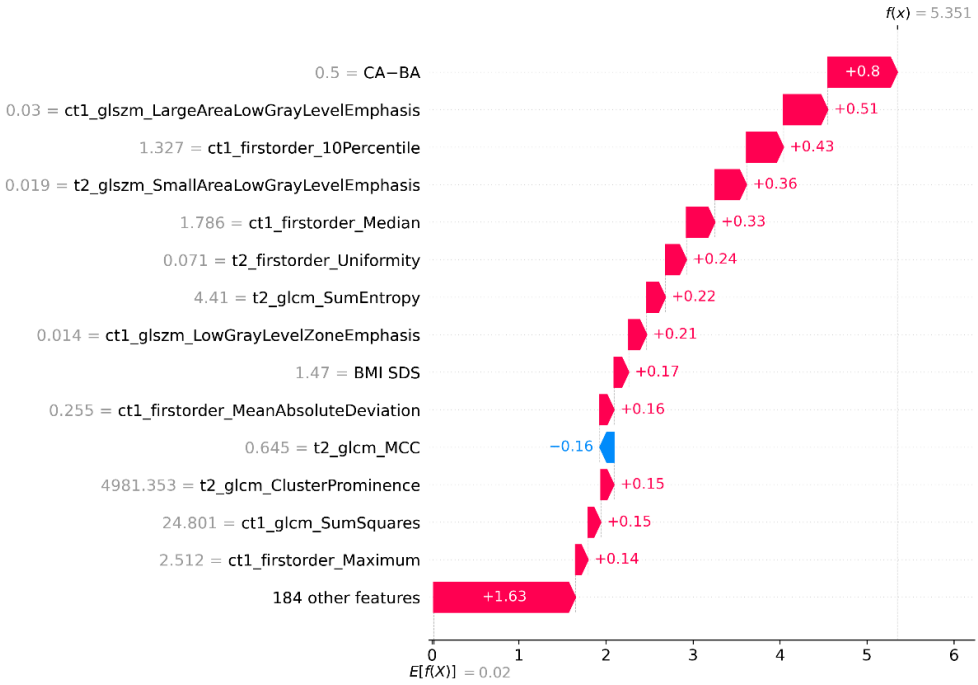

Figure 13. Representative waterfall plot of the combined model case.

Abbreviation: CA–BA; chronological age–bone age, GLSZM; gray-level size zone matrix, GLCM; gray-level co-occurrence matrix, BMI; body mass index; SDS; standard deviation score.

4. DISCUSSION

We demonstrated that the combined model using both clinical parameters and radiomics features accurately predicted GHD. The combined model was superior to the clinical and radiomics models. Among the clinical parameters, the BMI SDS, CA–BA, weight SDS, and growth velocity were the major contributing factors to the clinical model. Among the radiomics features, Inverse Variance from T2WI and Energy from T1C were the major factors contributing to the radiomics model. In the combined model, CA–BA, weighted SDS, Maximum Probability from T2WI, and Run Length Nonuniformity Normalized from T2WI were the major contributing factors.

Owing to the invasiveness and limitations of the GH provocation test, some studies have investigated prediction models using clinical parameters for GHD diagnosis. A single-center study from Argentina assessed clinical parameters including pituitary abnormalities, such as pituitary dysgenesis, midline abnormalities, and pituitary hormone deficiencies, in children and developed a GHD prediction model using a decision tree with internal validation only.[23] The sensitivity, specificity, and accuracy of the validation model were 55.6%, 99.2%, and 89.4%, respectively. A study from China developed a predictive model of GHD and ISS using clinical parameters, including IGF-1 and IGFBP-3, and MRI texture.[38] The AUC of the clinical and MRI texture predictive models were 0.607 and 0.852, respectively, although validation was not performed. Aim of our study was developing a clinical model for diagnosing GHD in children without pituitary abnormalities, systemic pathology, or endocrinological pathology, excluding GHD and ISS. We assessed various clinical parameters that can be easily obtained in local clinics and developed a machine learning model with external validation; the results were significant. Therefore, this model can be used to assess the etiology of short stature in real-world clinical settings.

To date, investigations of radiomics models for assessing children with short stature have been limited. Our previous study analyzed T2-weighted sella MRI images of children with short stature and developed a radiomics-based model for differentiation of GHD and ISS with internal validation, in which the AUC and accuracy were 0.705 and 70.6%, respectively.[16] In our study, the accuracy and AUC of the radiomics model were 0.698 and 0.691, respectively, for external validation. To improve the predictability of radiomics and clinical models, we combined both parameters using a machine learning classifier, XGBoost, to build the prediction models in this study. XGBoost is well-known for handling numerous features for model development with good performance, which is suitable for radiomics studies.[39,40] The pure radiomics model did not yield high predictive

performance in external validation; however, the combined clinical and radiomics model could accurately predict GHD with an AUC of 0.830 in external validation. Furthermore, the combined clinical and radiomics model yielded superior predictive performance compared with the clinical model. The added value of radiomics for predicting GHD was validated using an independent test set. Therefore, we believe that radiomics may have a predictive potential for differentiating between GHD and ISS.

To interpret the selected radiomic features and clinical parameters, we performed SHAP analysis. SHAP analysis enables quantification of the impact of radiomic features and clinical parameters on the prediction of GHD. SHAP estimates the importance value for each feature in the built model and facilitates informed clinical decision-making. We provided several SHAP plots to visualize the power of each selected feature on global (in the overall study population) and local (one patient) levels. This provides an intuitive visualization of how clinical and radiomic features contribute to the prediction of GHD. In both the radiomics and combined models, we found that the radiomic features extracted from both T1C and T2WI contributed to the prediction. Texture features and first-order features were used in the radiomics model. In the combined model, texture features were used for prediction. Shape features, including volume, were not used for the prediction, which is consistent with the fact that distinguishing GHD from ISS based on simple pituitary gland volume alone was not successful in previous studies. The Maximum Probability feature, a GLCM feature, was the most powerful predictor of GHD among the radiomic features, followed by the Run Length Non Uniformity Normalized, a GLRLM feature. GLCM measures the spatial distribution of gray-level intensities within an image, which is a biomarker for heterogeneity.[15] Particularly, as the Maximum Probability indicates occurrences of the most predominant pair of neighboring intensity values,[33] it may capture the different intensities of the pituitary gland between GHD and ISS, which cannot be detected by visual comparison. The GLRLM quantifies the gray-level runs, which are defined as the length of the number of pixels, of consecutive pixels that have the same gray-level value. A Run Length Non-Uniformity Normalized, one of GLRLM features, assess the homogeneity of run lengths, where higher similarity suggested by a lower value in run lengths throughout the image.[33] As higher values of Run Length Non-Uniformity Normalized showed significant association with GHD in our study, we can infer that more heterogeneous pituitary gland can be observed in GHD than in ISS.

Among the clinical parameters, BMI SDS, CA–BA, weight SDS, growth velocity, IGF-I SDS,

MPH SDS, and height SDS were the major contributing factors to the prediction model for GHD in our study. This result is consistent with those of previous studies. The clinical parameters related to the diagnosis of GHD have been investigated in several studies. IGF-1, a metabolic product influenced by GH, is synthesized in the liver and is mainly regulated by GH.[41] Thus, serum IGF-1 levels reflect the endogenous production of GH with minimal diurnal variation and have been considered as one of useful screening tests for GHD due to their diagnostic potential since 1982. In a meta-analysis, AUC of IGF-1 for diagnosis of GHD was 0.78.[42] A retrospective study reported that height velocity and IGF-1 could be used for screening of GHD.[43] A cohort study reported that BMI was negatively related with peak GH level on GH provocation test[22]. In addition, pubertal development is delayed in children with GHD which is related to bone age delay.[44,45] In a cohort study, bone age was more delayed in children with GHD compared to those with ISS.[26] In another cohort study, MPH was different according to etiologies of short stature.[2] Summary Statement of the Growth Hormone Research Society recommends to consider height SDS and height velocity when deciding whether to perform GH provocation test or not[1].

Our study has some limitations. Firstly, this study was limited to a single ethnicity with a retrospective design. Secondly, we could not consider IGFBP-3 because the values from both centers were significantly different owing to the different methods and reagents used. Thirdly, a genetic evaluation was not performed. Fourthly, the hypothalamus was not included in this analysis because sella MRI focus on pituitary glands. As the MRI protocol centers the field of view on the sella/suprasellar area, T2WI often fail to include the entire hypothalamus. In addition, the pituitary gland has relatively clear anatomical boundaries, making segmentation an easy task. However, the hypothalamus lacks clear anatomical boundaries, leading to difficulties in setting the region of interest. Consequently, the segmentation process itself is likely to be biased. MRI is still burdensome for children although it is less burdensome and GH provocation test which require multiple sampling and hospitalization. As sella MRI is performed for patients who has endocrinological problem, further studies investigate radiomics using various protocol of brain MRI is required for increment of practical value of radiomics for prediction of GHD and ISS.

However, this study is the first study which developed a prediction model for GHD using both clinical parameters including anthropometric data, laboratory test, and bone age and radiomics features. Moreover, we validated predictability of the model with external validation.

5. CONCLUSION

In conclusion, our research strongly emphasizes the potential of combining radiomics-based diagnostic models with clinical parameters for differentiation between GHD and ISS in children. This study meticulously analyzed both T2WI and T1C in sella MRI, alongside a comprehensive range of clinical parameters such as pubertal status and bone age, and scrutinized the individual contributions of these parameters to the predictive model. Our model combining both radiomics and clinical parameters can accurately predict GHD from ISS, which was also proved in the external validation, therefore proved its predictive potential. Subsequently, we may expect an individualized treatment strategy with our radiomics model combined with machine learning. Further studies with larger samples including various ethnicity and various brain MRI series are required to overcome limitation of our study. In addition, we hope to develop a robust model using genetic information as well as radiomics and clinical parameters to replace GH provocation test in the future.

REFERENCES

1. Consensus guidelines for the diagnosis and treatment of growth hormone (GH) deficiency in childhood and adolescence: summary statement of the GH Research Society. GH Research Society. J Clin Endocrinol Metab 2000;85:3990-3.
2. Song KC, Jin SL, Kwon AR, Chae HW, Ahn JM, Kim DH, et al. Etiologies and characteristics of children with chief complaint of short stature. Ann Pediatr Endocrinol Metab 2015;20:34-9.
3. Song K, Lee J, Lee S, Jeon S, Lee HS, Kim HS, et al. Height and subjective body image are associated with suicide ideation among Korean adolescents. Front Psychiatry 2023;14:1172940.
4. Voss LD. Short normal stature and psychosocial disadvantage: a critical review of the evidence. J Pediatr Endocrinol Metab 2001;14:701-11.
5. Hanew K, Utsumi A. The role of endogenous GHRH in arginine-, insulin-, clonidine- and l-dopa-induced GH release in normal subjects. Eur J Endocrinol 2002;146:197-202.
6. Martha PM, Jr., Gorman KM, Blizzard RM, Rogol AD, Veldhuis JD. Endogenous growth hormone secretion and clearance rates in normal boys, as determined by deconvolution analysis: relationship to age, pubertal status, and body mass. J Clin Endocrinol Metab 1992;74:336-44.
7. Bidlingmaier M. Problems with GH assays and strategies toward standardization. Eur J Endocrinol 2008;159 Suppl 1:S41-4.
8. Ariza Jiménez AB, Martínez Aedo Ollero MJ, López Siguero JP. Differences between patients with isolated GH deficiency based on findings in brain magnetic resonance imaging. Endocrinol Diabetes Nutr (Engl Ed) 2020;67:78-88.
9. Kessler M, Tenner M, Frey M, Noto R. Pituitary volume in children with growth hormone deficiency, idiopathic short stature and controls. J Pediatr Endocrinol Metab 2016;29:1195-200.
10. Nagel BH, Palmbach M, Petersen D, Ranke MB. Magnetic resonance images of 91 children with different causes of short stature: pituitary size reflects growth hormone secretion. Eur J Pediatr 1997;156:758-63.
11. Oh JS, Sohn B, Choi Y, Song K, Suh J, Kwon A, et al. The influence of pituitary volume

on the growth response in growth hormone-treated children with growth hormone deficiency and idiopathic short stature. Ann Pediatr Endocrinol Metab 2023; doi:10.6065/apem.2346052.026.

12. Wagner MW, Bilbily A, Beheshti M, Shammas A, Vali R. Artificial intelligence and radiomics in pediatric molecular imaging. Methods 2021;188:37-43.
13. Madhogarhia R, Haldar D, Bagheri S, Familiar A, Anderson H, Arif S, et al. Radiomics and radiogenomics in pediatric neuro-oncology: A review. Neurooncol Adv 2022;4:vdac083.
14. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-77.
15. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev 2016;1:207-26.
16. Lee T, Song K, Sohn B, Eom J, Ahn SS, Kim HS, et al. A Radiomics-Based Model with the Potential to Differentiate Growth Hormone Deficiency and Idiopathic Short Stature on Sella MRI. Yonsei Med J 2022;63:856-63.
17. Huang J, Shlobin NA, Lam SK, DeCuypere M. Artificial Intelligence Applications in Pediatric Brain Tumor Imaging: A Systematic Review. World Neurosurg 2022;157:99-105.
18. Suh J, Heo J, Kim SJ, Park S, Jung MK, Choi HS, et al. Bone Age Estimation and Prediction of Final Adult Height Using Deep Learning. Yonsei Med J 2023;64:679-86.
19. Rui W, Wu Y, Ma Z, Wang Y, Wang Y, Xu X, et al. MR textural analysis on contrast enhanced 3D-SPACE images in assessment of consistency of pituitary macroadenoma. Eur J Radiol 2019;110:219-24.
20. Zhang Y, Chen C, Tian Z, Cheng Y, Xu J. Differentiation of Pituitary Adenoma from Rathke Cleft Cyst: Combining MR Image Features with Texture Features. Contrast Media Mol Imaging 2019;2019:6584636.
21. Stanley T. Diagnosis of growth hormone deficiency in childhood. Curr Opin Endocrinol Diabetes Obes 2012;19:47-52.
22. Yang A, Cho SY, Kwak MJ, Kim SJ, Park SW, Jin DK, et al. Impact of BMI on peak growth hormone responses to provocative tests and therapeutic outcome in children with growth hormone deficiency. Sci Rep 2019;9:16181.
23. Clément F, Grinspon RP, Yankelevich D, Martín Benítez S, De La Ossa Salgado MC, Ropelato MG, et al. Development and Validation of a Prediction Rule for Growth Hormone

Deficiency Without Need for Pharmacological Stimulation Tests in Children With Risk Factors. Front Endocrinol (Lausanne) 2020;11:624684.

24. Song K, Jung MK, Oh JS, Kim SJ, Choi HS, Lee M, et al. Comparison of growth response and adverse reaction according to growth hormone dosing strategy for children with short stature: LG Growth Study. Growth Horm IGF Res 2023;69-70:101531.
25. Kim JH, Yun S, Hwang SS, Shim JO, Chae HW, Lee YJ, et al. The 2017 Korean National Growth Charts for children and adolescents: development, improvement, and prospects. Korean J Pediatr 2018;61:135-49.
26. Yoon JY, Cheon CK, Lee JH, Kwak MJ, Kim HJ, Kim YJ, et al. Response to growth hormone according to provocation test results in idiopathic short stature and idiopathic growth hormone deficiency. Ann Pediatr Endocrinol Metab 2022;27:37-43.
27. Ranke MB, Lindberg A. Observed and predicted growth responses in prepubertal children with growth disorders: guidance of growth hormone treatment by empirical variables. J Clin Endocrinol Metab 2010;95:1229-37.
28. Huynh QTV, Ho BT, Le NQK, Trinh TH, Lam LHT, Nguyen NTK, et al. Pathological brain lesions in girls with central precocious puberty at initial diagnosis in Southern Vietnam. Ann Pediatr Endocrinol Metab 2022;27:105-12.
29. Chotipakornkul N, Onsoi W, Numsriskulrat N, Aroonparkmongkol S, Supornsilchai V, Srilanchakon K. The utilization of basal luteinizing hormone in combination with the basal luteinizing hormone and follicle-stimulating hormone ratio as a diagnostic tool for central precocious puberty in girls. Ann Pediatr Endocrinol Metab 2023;28:138-43.
30. Hyun SE, Lee BC, Suh BK, Chung SC, Ko CW, Kim HS, et al. Reference values for serum levels of insulin-like growth factor-I and insulin-like growth factor binding protein-3 in Korean children and adolescents. Clin Biochem 2012;45:16-21.
31. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist: Stanford university press; 1959.
32. Avants B, Tustison N, Song G. Advanced Normalization Tools (ANTS) Insight J. 2009: 1–35.
33. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-e7.

34. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996;58:267-88.
35. Du N, Song L, Gomez-Rodriguez M, Zha H. Scalable Influence Estimation in Continuous-Time Diffusion Networks. Adv Neural Inf Process Syst 2013;26:3147-55.
36. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p.785-94.
37. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
38. Cong M, Qiu S, Li R, Sun H, Cong L, Hou Z. Development of a predictive model of growth hormone deficiency and idiopathic short stature in children. Exp Ther Med 2021;21:494.
39. Nazari M, Shiri I, Zaidi H. Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients. Comput Biol Med 2021;129:104135.
40. Wang X, You X, Zhang L, Huang D, Aramini B, Shabaturov L, et al. A radiomics model combined with XGBoost may improve the accuracy of distinguishing between mediastinal cysts and tumors: a multicenter validation analysis. Ann Transl Med 2021;9:1737.
41. Hadjadj S, Faure-Gerard C, Ragot S, Millet C, Duengler F, Torremocha F, et al. Diagnostic strategy for growth hormone deficiency: relevance of IGF-1 determination as a screening test. Ann Endocrinol (Paris) 2007;68:449-55.
42. Shen Y, Zhang J, Zhao Y, Yan Y, Liu Y, Cai J. Diagnostic value of serum IGF-1 and IGFBP-3 in growth hormone deficiency: a systematic review with meta-analysis. Eur J Pediatr 2015;174:419-27.
43. Lemaire P, Brauner N, Hammer P, Trivin C, Souberbielle JC, Brauner R. Improved screening for growth hormone deficiency using logical analysis data. Med Sci Monit 2009;15:Mt5-10.
44. Zadik Z, Chalew S, Zung A, Landau H, Leiberman E, Koren R, et al. Effect of long-term growth hormone therapy on bone age and pubertal maturation in boys with and without classic growth hormone deficiency. J Pediatr 1994;125:189-95.
45. Rhie YJ, Yoo JH, Choi JH, Chae HW, Kim JH, Chung S, et al. Long-term safety and effectiveness of growth hormone therapy in Korean children with growth disorders: 5-year

results of LG Growth Study. PLoS One 2019;14:e0216927.

ABSTRACT (IN KOREAN)

Sella 자기공명영상 기반 라디오믹스 및 임상적 지표를 이용한 성장호르몬결핍증 과 특발성 저신장의 진단 모델 개발 및 검증

성장호르몬 결핍증과 특발성 저신장의 감별 진단을 위한 검사로 성장호르몬 유발 검사가 gold standard로 알려져 있지만 매우 침습적이며 타당성과 재현성의 한계점을 가지고 있다. 따라서, 성장호르몬 결핍증과 특발성 저신장의 감별진단을 위한 예측 모델에 대한 연구가 필요한 실정이다. 라디오믹스는 수학적 알고리즘을 사용하여 다양한 기능을 추출하는 방법으로, 인간 눈으로 감지할 수 없는 분자 프로필과 질병 특성을 찾을 수 있다. 생체 의학 이미지의 정보 개념을 기반으로, 라디오믹스는 디지털 의료 이미지를 추출 가능한 고차원 데이터로 변환한다. 그러나 아직 소아 뇌하수체에 대한 라디오믹스 연구는 제한적이다. 따라서 우리는 sella 자기 공명 영상 및 임상 변수를 포함한 라디오믹스를 사용한 성장호르몬 결핍증 및 특발성 저신장 진단의 머신 러닝 기반 예측 모델을 개발하고 외부 검증을 통해 이 모델의 타당성을 검증하고자 했다.

독립된 병원에서 정상 sella 자기공명영상 결과를 가진 소아 293명으로 구성된 training set와 47명의 test set로 연구를 진행했다. Sella 자기공명영상에서는 T2 weighted image 및 contrast-enhanced T1-image의 뇌하수체에서 총 186개의 라디오믹스 변수를 추출했다. 임상 변수에는 계측 지표, insulin-like growth factor-I (IGF-I) 및 골 연령이 포함되었다. XGBoost 알고리즘을 사용하여 예측 모델을 훈련시켰다. Internal validation은 training set에서 five-fold cross validation으로 수행했고, external validation은 test set에서 수행했다. 모델의 성능은 area under the receiver operating characteristic curve (AUC)을 통해 평가했다. 각 변수의 영향을 정량화하기 위해 mean absolute Shapley values을 계산하였다.

External validation에서 임상, 라디오믹스 및 병합 모델의 AUC는 각각 0.684, 0.691 및 0.830이었다. 임상 모델에서 예측에 대한 주요 기여 요인은 체질량 지수 표준 편차, 역연령과 골 연령의 차이, 체중 표준 편차, 성장 속도 및 IGF-I 표준 편차였다. 라디오믹스 중에서는 T2 weighted image의 Inverse Variance 및 contrast-enhanced T1-image의 energy가 라디오믹스 모델에 기여하는 주요 요인이었다. 병합 모델에서는 라디오믹스 기능이 예측력을 강화했다.

이 연구는 성장호르몬 결핍증과 특발성 저신장의 전통적인 진단 기법의 한계를 극복하기 위한 라디오믹스 기반 진단 모델의 잠재력을 보여주었다. 이러한 결과는 소아과학에서 라디오믹스와 머신 러닝의 중요한 역할을 입증하며 이를 이용한 새로운 진단 기법 발전의 가능성을 뒷받침한다.

핵심되는 말 : 성장호르몬 결핍증, 특발성 저신장, 머신러닝, 라디오믹스