Early identification and diagnosis of growth disorders using natural language processing and machine learning

Chun-Yen Cheng[1], Nguyen Quoc Khanh Le[1,*], Yung-Chun Chang[2,*], Jason C. Hsu[2], Chao-Hsu Lin[3,4], Chen Yang[5,6], Chung-Ming Chen[5,6], Jia-Woei Hou[7], Min-Huei Hsu[2,*]

[1]Professional Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]Graduate Institute of Data Science, Taipei Medical University, Taiwan.
[3]Department of Pediatrics, Hsinchu Mackay Memorial Hospital, Hsinchu, Taiwan
[4]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[5]Department of Pediatrics, Taipei Medical University Hospital, Taipei, Taiwan
[6]Department of Pediatrics, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[7]Department of Pediatrics, Cathay General Hospital, Taipei, Taiwan

Background

Growth abnormality is an essential diagnosis in the clinical practice of pediatricians, and different growth disorder forms should be detected, diagnosed, and treated early by routine monitoring. A previous study showed that using Electronic Health Records (EHRs) to diagnose growth disorders in children with an automated growth monitoring system significantly improved the diagnosis rate and growth-related referrals to pediatricians, and the average delay in diagnosis improved from partially over 5 to 1.97 years.

However, the standardization of testing tools was deficient with the specialist' Electronic Medical records (EMRs), and the evidence supporting growth monitoring is suboptimal in most countries, still leading to delays in growth diagnosis, lack of appropriate testing tools, and inappropriate referrals.

➔ This study aims to use EMRs and machine learning to build an electronic growth chart model to assist primary care physicians in interpreting growth disorders in Taiwanese children.

Methods

This retrospective study used the Taipei Medical University Clinical Research Database (TMUCRD) with EMR data from two hospitals. The data were mapped to OHDSI OMOP CDM, which is expected to be developed into multinational cooperative research. The study used a total of 111,214 clinical growth data from 15,627 outpatients for the five years from January 2016 to December 2020 for further analysis.

Text mining (using AI with natural language processing (NLP)) and data preprocessing have been applied to extract and clean unstructured data. The research aims to focus on the information of new outpatients during the five years and tracked in outpatient clinics for more than one year. 20,807 data were from Taipei Medical University Hospital (TMUH) as the training, testing, and internal validation sets. Moreover, 4,212 data from WanFang Hospital (WFH) served as an external validation set.

This study implemented different machine learning algorithms to predict growth disorders in outpatients after one year of follow-up. We assessed the performance of each model using various measurement metrics to find the optimal one. Feature selection and imbalance approaches are employed to find the optimal feature set with a balanced result. Furthermore, we analyzed the electronic growth chart of Taiwanese children based on the model's results. The final results showed that we could track the standard deviation of target height ≥1 or ≥2 SDS, skeletal age value and chronological age ≥1 or ≥2 SDS, height percentile and weights percentile, and growth rate ≤5 cm/year to improve the diagnosis of growth disorders.

Contact: G167109007@tmu.edu.tw

Results

The text mining of EMR extracts 12 features: mid-parental height, the parents' height, height records, weight records, skeletal age, chronological age, X-ray date, primary diagnosis, all diagnoses, and drugs. It also combined a growth dataset of Taiwanese children to extract 11 features, including difference and standard deviation of skeletal age value, percentile interval and median percentile of mid-parental height and heights and weights, annual growth rate, and standard deviation of target height. Then, the data processing results showed no differences between traditional text mining and NLP methods (overall matching accuracy from 89.19% to 99.6%).

Analyzing essential features for all new outpatients under the first 12 times records module or hybrid feature selection module (54 features) analysis, all datasets performed well and consistently among RF, GBM, and XGB models. RF algorithm got the high performance of short stature or precocious puberty in classification and diagnosis; and with a faster computational time. It achieved an accuracy of 0.88, sensitivity of 0.91, specificity of 0.86, F1-score of 0.88, and AUC of 0.89. In addition, the performance of the GH2 classification and verification of the skeletal age value ≧2 SDS, or target height≧2 SDS or growth rate≦5 cm/year, is more significant and excellent, with an accuracy of 0.90, sensitivity of 0.92, specificity of 0.87, F1-score of 0.91, and AUC of 0.89.

Table 1. Patients' demographic

	Overall Patient	TMUH Data	TMUH Patient	WFH Data	WFH Patient
Overall	15627	85743	11614	25471	4013
Girls	8614	47346	6280 (54.1%)	15226	2334 (58.2%)
Boys	7013	38394	5334 (45.9%)	10245	1679 (41.8%)
Follow-up patients	9764	57711	6998 (60.3%)	19140	2766 (68.9%)
New Patients	5863	28029	4616 (39.7%)	6331	1247 (31.1%)
New patient girls	3274	16197	2519 (54.6%)	4021	755 (60.5%)
Newly patient boys	2189	11836	2097 (45.4%)	2310	492 (39.5%)
New patient for follow-up 1year	2129	20807	1739	4212	390
New patient for follow-up 1year girls	1221	12231	974 (56.0%)	2748	247 (63.3%)
New patient for follow-up 1year boys	908	8576	765 (44.0%)	1464	143 (36.7%)
Specialist 1		47810		25471	
Specialist 2		37933		0	

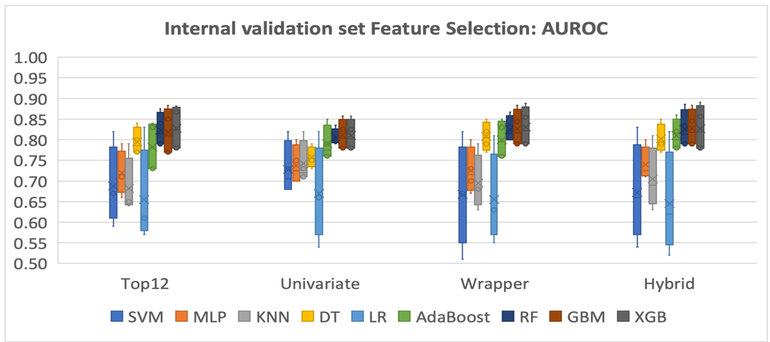

Figure 1. Performance of Prediction Models under various Feature Selection Models

Table 2. Performance of Prediction Models

TOP12	Internal validation: Diagnosis (TMUH)							External validation: Diagnosis (WFH)						
	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC
RF	0.89	0.81	0.86	0.82	0.84	0.84	0.84	0.87	0.84	0.91	0.85	0.87	0.88	0.88
GBM	0.88	0.83	0.85	0.84	0.84	0.85	0.85	0.88	0.84	0.89	0.85	0.86	0.87	0.87
XGB	0.88	0.84	0.89	0.84	0.86	0.87	0.87	0.87	0.83	0.89	0.84	0.86	0.86	0.87
Hybrid	**Internal validation: Diagnosis**							**External validation: Diagnosis**						
	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC
RF	0.87	0.81	0.85	0.82	0.83	0.83	0.83	0.86	0.85	0.91	0.86	0.88	0.88	0.89
GBM	0.88	0.82	0.86	0.83	0.84	0.85	0.85	0.87	0.84	0.89	0.86	0.87	0.88	0.88
XGB	0.88	0.83	0.88	0.85	0.83	0.86	0.86	0.87	0.85	0.88	0.86	0.86	0.87	0.87

TOP12	Internal validation: GH2							External validation: GH2						
	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC
RF	0.88	0.91	0.89	0.86	0.9	0.88	0.88	0.88	0.9	0.92	0.86	0.91	0.9	0.89
GBM	0.87	0.92	0.89	0.88	0.9	0.88	0.88	0.88	0.9	0.9	0.86	0.9	0.88	0.88
XGB	0.88	0.92	0.88	0.9	0.9	0.88	0.88	0.89	0.9	0.89	0.87	0.9	0.88	0.88
Hybrid	**Internal validation: GH2**							**External validation: GH2**						
	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC	cross_val_acc	precision	sensitivity	specificity	f1 score	accuracy	AUC
RF	0.87	0.93	0.88	0.89	0.90	0.88	0.89	0.88	0.91	0.92	0.87	0.91	0.90	0.89
GBM	0.87	0.92	0.89	0.88	0.90	0.88	0.88	0.89	0.90	0.90	0.86	0.90	0.89	0.88
XGB	0.88	0.92	0.90	0.89	0.91	0.89	0.89	0.88	0.90	0.89	0.87	0.90	0.88	0.88

Conclusions

This study uses different machine learning algorithms, which have stable and excellent performance in classifying and diagnosing growth disorders in all newly diagnosed children. RF is a convenient algorithm for accurate medical diagnosis among all the above algorithms. They helped improve the accuracy and served as the model for the electronic growth chart. As an application, our machine learning model can be used in identifying the future visual system of short stature using our signature features.