IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2017, accepted April 18, 2017, date of publication April 24, 2017, date of current version June 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2696704

A Data-Driven Decision Support System for Scoliosis Prognosis

LIMING DENG[1], YONG HU[2], (Senior Member, IEEE), JASON PUI YIN CHEUNG[2], AND KEITH DIP KEI LUK[2]

[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong
[2]Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong

Corresponding author: Liming Deng (limngdeng2-c@my.cityu.edu.hk)

This work was supported by the Hong Kong Health and Medical Research Fund under Grant 03142306.

ABSTRACT A decision support system with data-driven methods is of great significance for the prognosis of scoliosis. However, developing an accurate and interpretable data-driven decision support system is challenging: 1) the scoliosis data collected from clinical environments is heterogeneous, unstructured, and incomplete; 2) the cause of adolescent idiopathic scoliosis is still unknown, and the effects of some measured indicators are not clear; and 3) some treatments like wearing a brace will affect the progression of scoliosis. The main contributions of the paper include: 1) propose and incorporate different imputation methods like Local Linear Interpolation (LLI) and Global Statistic Approximation (GSA) to deal with complicated types of incomplete data in clinical environments; 2) identify important features that are relevant to the severity of scoliosis with embedded method; and 3) establish and compare the scoliosis prediction models with multiple linear regression, k nearest neighbor, tree, support vector machine, and random forest algorithms. The prediction performance is evaluated in terms of mean absolute error, root mean square error, mean absolute percentage error, and the Pearson correlation coefficient. With only a few critical features, the prediction models can achieve satisfactory performance. Experiments show that the models are highly interpretable and viable to support the decision-making in clinical environments.

INDEX TERMS Scoliosis prognosis, missing values, feature selection, decision support system, data-driven method.

I. INTRODUCTION

Idiopathic scoliosis is a complex spine distortion that affects millions of people in the world [1]. The commonly occurring symptoms of scoliosis include poor personal image, poor truncal balance, and susceptibility to back pain. A large spinal deformity may cause cardiopulmonary compromise resulting in significant health complications or even death [2], [3]. The severity of scoliosis is usually quantified by Cobb angle, which is the angle between the two most tilted vertebrae in the spinal curves [4]. Treatment of scoliosis depends on the magnitude of Cobb angle and whether it is progressing. Most scoliosis patients present with minor curves, which requires periodical examinations to determine any progression. Those with moderate curves or progressive curves may be treated with bracing. Only about 0.25% scoliosis cases will require surgery [5]. A prediction of the curve progression can help doctors determine which treatment method is most appropriate. However, it is still a challenge for physicians to determine which scoliosis patients may progress and most rely on their experience and foresight.

A few studies [6]–[8] have attempted to deal with these challenges by predicting the Cobb angle or its progression with data-driven methods. The data-driven methods can provide reasonable support for doctors to address these two problems: 1) whether the patient needs close observations in future; 2) whether the patient should be treated preemptively in anticipation of future progression.

Wu *et al.* [3] combined fuzzy c-means clustering methods and artificial neural network (ANN) to predict the follow-up scoliosis Cobb angle. The material used included 61 scoliosis patients with at least four follow-up Cobb angle measurements. This method clusters scoliosis patients into several groups and assumes their progression follow different patterns [9]. The model may not be applicable for those scoliosis patients with few historical records as well as new scoliosis patients. Chalmers *et al.* [4], [10] proposed a conditional fuzzy clustering model for predicting the progression of braced scoliosis. Although the fuzzy method can help the clustering, the importance of features and their interactions may not be well identified. Ajemba *et al.* [11] utilized

common indicators of patients to predict the progression of scoliosis with support vector machine. However, the method may find it difficult to apply to those patients with missing key values.

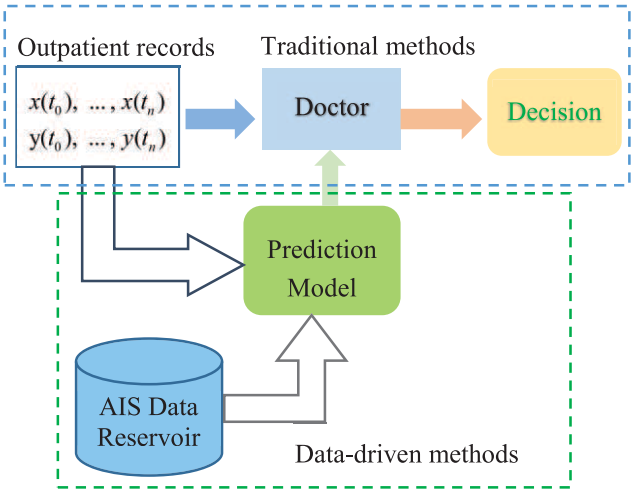


FIGURE 1. Decision support system for scoliosis prognosis.

Based on the above analysis, we propose a decision support system for scoliosis prognosis, as shown in Fig. 1. Our study aims at predicting the scoliosis curvature of patients for clinical treatments. The initial results of this work has been reported [12]. In this paper, the situations of missing values in clinical environments have been intensively analyzed and handled. The importance of features has been identified and selected for the prediction models.

II. PROBLEM DESCRIPTION

In clinical practice, the treatment decisions made by doctors for scoliosis patients usually can be decomposed into three steps:

Step 1: Diagnosis: Check the symptoms (current diagnosis records and historical diagnosis records) of the patient;

Step 2: Prognosis: Identify or estimate the curvature of the spine in the future;

Step 3: Decision-making: Find appropriate therapies for the patient based on the diagnosis and prognosis.

There exist large uncertainties during this decision-making process, especially in the prognosis step, which can be affected by many uncontrollable factors such as the experience or skill level of attending doctors. Data-driven methods can provide quantitative support for physicians. The decision support system utilizes both the current outpatient diagnosis records and the other data in the Scoliosis Data Reservoir to build this prediction model, as shown in Fig. 1, which can reduce the uncertainties of traditional decision-making.

However, there are two main challenges in building the prediction model for scoliosis in the clinical environment.

1) The scoliosis data are incomplete, and the sources of the incomplete data are very complicated. Some records are unavailable due to the following reasons:
   - Some tests may not be available for the lack of the necessary equipment;
   - Some measurements may not be appropriate for certain patients;
   - Different doctors may require different measurements.
2) Although there are many risk factors [4], [11], [13]–[15] have been reported that relate to the progression of scoliosis, it is hard to identify critical ones and quantify the effects because of the unknown causes [16].

In general, the incomplete data from complicated sources and the unclear relation of the indicators are two major obstacles to having a reliable prediction model for clinical treatment. In addition, the interpretability is another important feature that the prediction model should possess.

III. MATERIAL AND METHOD

A. OVERVIEW

When a new patient comes to seek advice or treatments for his/her scoliosis problem, the individual data of the patient is not enough to build a reliable data-driven model for the decision support. The data from other patients would be a good reference for the decision support. This idea motivates the study to establish a reliable and interpreted prediction model by incorporating the records both from the new patient and other patients. However, the scoliosis data is often unstructured data with many missing values, missing features or even text description data, which restrict the available data samples for the prediction model. The main work of this paper is to reduce the complexity of the scoliosis data, so as to be understood and modeled easily for prognosticating. The overall method is shown in Fig. 2, and the methods are outlined in the followings:

- The scoliosis data are recorded as multivariate longitudinal data with much unaccountable noise, missing values, and categorical features. The clinical data should be preprocessed to reduce the noise, impute the missing values and transfer the categorical features to numeric features.
- Based on the preprocessed data, new features are extracted from the historical records, the importance of features are evaluated with random forest importance metric and Pearson correlation coefficients for the selection of best feature subset. The features are added to the initial feature subset sequentially according to their importance, so as to avoid the exhaustive search. 5-fold cross-validation is employed to select the best feature subset.
- The prediction models are established with multivariate linear model, k nearest neighbor regression, tree regression, support vector machine and random forest

TABLE 1. Summary of the recorded data.

Indicators	Description	Values and range	Missing value (%)
ID No.	Indicate the identity of each patient	Integer	No
Gender	Male or female	Category {F, M}	<1%
Date of birth	Birthday	Time stamp	2%
FamHx	Indicate whether family members have scoliosis	Category {Y, N, -}	80%
Appt Date	Appointment date for the diagnosis	Time stamp	5%
Brace	Indicate the treatment for the scoliosis	Category {+, -, missing}	35%
BodyHeight	Body height (cm)	Numeric {79∼189.2}	17%
SittingHeight	Sitting height (cm)	Numeric {49∼102}	68%
ArmSpan	Arm span (cm)	Numeric {55∼190}	24%
RisserSign	Describe the maturity of bone	Numeric {0∼5}	27%
CobbStanding	Reflect the curvature location	Category, like T2/T9	36%
CobbAngle	Indicate the magnitude of the curvature (degree)	Numeric { 0∼118}	8%
Follow up interval	The follow up time period between two adjacent appointment date	Appt Date 2 - Appt Date 1	>50%
Radius	Indicate the maturity for the radius bone	{0,1,...,11}	>50%
Ulna	Indicate the maturity for the ulna bone	{0,1,...,11}	>50%
Phalanx	Indicate the maturity for the phalanx bone	Numeric	>50%
Menarche	Time stamp for menarche	Time stamp	>50%
Age.Menarche.years	The age at menarche for Female only	Age (years)	>50%
Change of CobbAngle	The change of the Cobb angle compared to last record	Float data	>50%
SRS.score	The quality of life score assessed with instrument developsed by SRS	Numeric	>50%

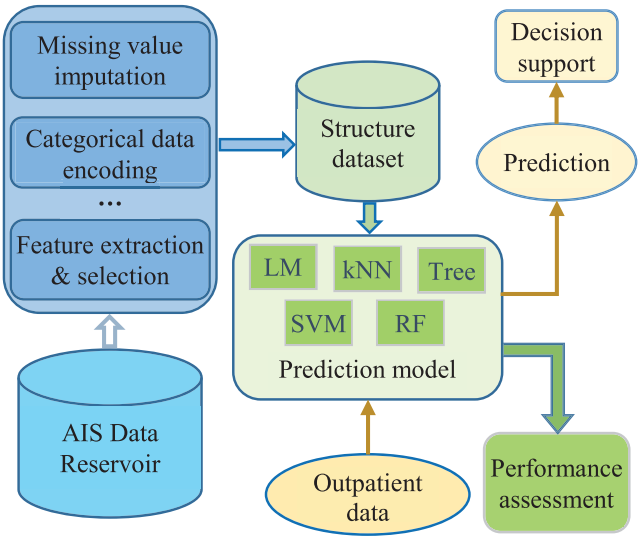


FIGURE 2. The configuration of data-driven methods for scoliosis prognosis.

method. The cross-validation that leaves one patient out is employed to compare their prediction performance.

B. SCOLIOSIS DATA DESCRIPTION

Our data included scoliosis medical records of patients from the Duchess of Kent Children's Hospital in Hong Kong. More than two thousand patients have been involved in this project. About twenty indicators have been recorded by the hospital, and Table 1 summarizes the recorded indicators. The Cobb angle indicates the severity of scoliosis. The larger the Cobb angle, the worse the scoliosis. The other indicators reflecting the conditions of patients are assumed to be related with scoliosis. Some indicators from clinical environments are measured and recorded manually with unstructured properties, and hard to be utilized. The unstructured characteristics of the collected data are summarized in the following aspects:

i Some records of the indicators are missing, and the percent of the missing records are shown in Table 1. The indicators with many missing values are incapable of providing useful information when compared to the noise or measurement error;

ii The indicators, like *Brace*, *Gender* or *FamHx*, with categorical attribute cannot be fed into data-driven model directly;

iii The number of spine curvatures may be variable. About half of the patients have two or more curves. Besides, the progression patterns of curves are also varied from patient to patient.

The above phenomena make data extraction difficult for the prediction of scoliosis. Only part of the indicators have been considered in the prediction models for its simplicity and representativeness. For example, some indicators like *BodyHeight*, *SittingHeight* and *ArmSpan* are highly correlated, thus, only *BodyHeight* has been incorporated into our models. Furthermore, the progression process for double or triple curves is more complicated than the single one. To simplify the problem, the study concentrates on the prediction of single curve cases. The paper analyzed total 341 patients with 80 males and 261 females. The number of records for each patient is no less than two. The diagnosis age for most patients is from 10 to 20 years old. The time span for predicting the next *CobbAngle* is most within one year. Most of the *CobbAngle* lie within 10° to 50°. The histogram distribution for *currentAge*, *TimeSpan* and *currentCobbAngle* as well as *futureCobbAngle* are shown in Fig. 3.

The categorical features are encoded with the dummy variable method [17], and the encoded values are summarized in Table 2. Some features are extracted from the primary indicators to be more representative, such as *Age* extracted from *Birthday* and *Appointment Date*. The *TimeSpan* extracted as the period between two adjacent age parameters for each

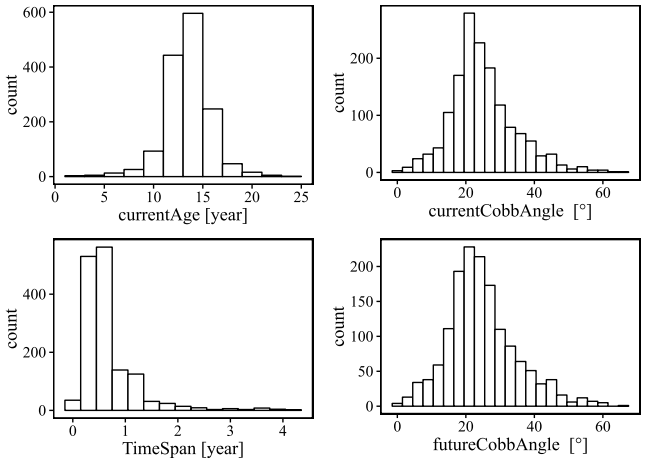


FIGURE 3. The histogram distribution for *currentAge*, *TimeSpan* and *currentCobbAngle* as well as *futureCobbAngle*.

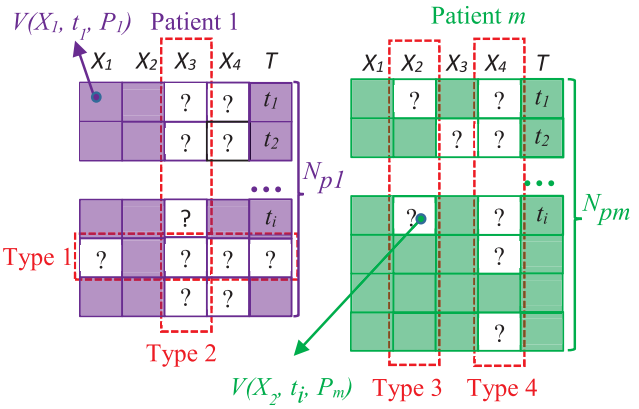


FIGURE 4. The main types of missing values for the scoliosis data (X_1, X_2, X_3, X_4 and T represent some features of patients; here, T represents the feature of appointing time and also can be converted to age; t_1, t_2 or t_i represent the specific time or age and V represent the value of the corresponding feature at that time for each patient).

TABLE 2. Encoding categorical features.

Features	*Gender*		*FamHx*			*Brace*		
Categorical	F	M	Y	N	missed	+	-	missed
Numeric	1	0	1	-1	0	1	-1	0

TABLE 3. Newly constructed features.

Extracted Features	*currentCobbAngle*, *futureCobbAngle*, *currentAge*, *TimeSpan*, *ccurrentBrace*, *FutureBrace*, *changeBrace*

patient. The *currentBrace*, *futureBrace* and *changeBrace* are extracted from the Brace states. Brace wearing was determined by the doctors and is prospective, thus, the future information of brace statement can be incorporated into the prediction model. Since our prediction model aims at predicting the next magnitude of Cobb angle, the Cobb angle is divided into *currentCobbAngle* and *futureCobbAngle*. The *currentCobbAngle* indicates the initial value of Cobb angle. The *futureCobbAngle* indicates the next Cobb angle of the same patient, which is selected as the responsible variable. The extracted features are summarized in Table 3.

C. MISSING VALUE IMPUTATION

The missing values are common and a tough problem to solve in clinical datasets. The reasons varied from the manually missing entry process, equipment errors to incorrect measurements [18]. There are many methods for dealing with missing values. Some simple methods utilize the most common values such as mean or median to replace the missing values, and some machine learning methods like k nearest neighbor or random forest methods to impute the missing values [19], [20]. However, handling the missing values is very specific and domain knowledge related. None of the solutions above can really directly solve the problem of missing values.

In our preprocessing procedure, dealing with missing values is one of the most challenging tasks. The situation of missing values in our dataset is very complicated. Fig. 4 indicates the types of missing values, T represents each time the patient visits the clinic for assessment. X_1,X_2, X_3 and X_4 represent part of the measured indicators for each patient. The "?" indicates the missing value. The main types of missing values are defined for easy understanding and preprocessing.

Type 1: one appointment record is missing most of the indicators;

Type 2: one indicator is missing in most of the records for an individual patient, while other patients have almost complete records for this indicator;

Type 3: the indicator is missing in only a few or part of the records within an individual patient;

Type 4: the indicator is missing in most of the records for most of the patients.

A single imputation method is incapable of handling this complicated situation of missing values. We combine domain knowledge to analyze the properties of each feature and then handle the four types of missing values accordingly. Case deletion is utilized in Type 1 and Type 4, because the instances and the features with many missing values may bring larger noise than the useful information they would bring to the prediction model; As for Type 2 and Type 3 missing conditions, if the indicators are categorical (like *FamHx*, *Brace*), the missing values are considered as a new category and encoded with a new value;

As for the numeric features (like *BodyHeight*, *RisserSign*), these features are assumed to be correlated with age ignoring the measurement errors and other noise. The Local Linear Interpolation (LLI) and Global Statistic Approximation (GSA) methods are proposed for the imputation of Type 3 and Type 2 conditions respectively.

a) The LLI method assumes that the feature of a patient is linearly changed within a short growth period, and the

missing values are imputed with interpolation method. The LLI method is dealing with the Type 3 missing condition, like features X_2, X_3 of patient m in Fig. 4. The Formula (1) shows the implementation of LLI method;

$$V(X_k, t_m) = V(X_k, t_i) + (V(X_k, t_j) - V(X_k, t_i))\frac{t_m - t_i}{t_j - t_i} \quad (1)$$

where, $V(X_k, t_m)$ is the missing value of an individual patient on feature X_k at appointed time (or age) t_m, $V(X_k, t_j)$ and $V(X_k, t_i)$ are the time closest values for the same indicator. The value (V) of the missing part and the measured records are not discriminated with the patient because all the V used in Formula (1) are from the same patient. t_m, t_i and t_j represent the appointed time (or age) when the features are measured.

b) For the missing condition of Type 2, the feature of an individual patient is absent for almost all the values, and the LLI method is inappropriate for the feature because of lacking enough information for the interpolation, the GSA method can be utilized to approximate the most likely value from other patients on similar growth period. Formula (2) shows the implementation of GSA.

$$V(X_k, t_m, P_i) = \frac{1}{n}\sum_{j \neq i} V(X_k, t_q, P_j) \quad |t_q - t_m| < \gamma \quad (2)$$

where, $V(X_k, t_m, P_i)$ is the missing value for the i-th patient on feature X_k at age t_m, $V(X_k, t_q, P_j)$ represents the value for j-th patient on feature X_k at age t_q that similar to the age t_m. The γ controls the similiarity of the age. Formula (2) utilizes other patients' non-missing values for the imputation.

By combining the above methods, the existed missing value problem can be well handled, and the procedures are summarized in the following steps:

Step 1: Detecting missing values from all the instances (data row in Fig. 4), and the instances are deleted if many features are absent.

Step 2: detecting the missing values over all the features (data column in Fig. 4), if a feature misses most of the values, then delete the feature;

Step 3: if a categorical variable is detected with missing values, then encode a new value to these missing values, as shown in Table 2;

Step 4: if a numerical variable is detected with missing values, then apply LLI method to those patients with enough individual features to support the imputation, the other patients who failed to implement the LLI method can utilize the GSA method for the imputation.

D. FEATURE SELECTION

The aim of feature selection is to improve the prediction accuracy, reduce the computation cost and a better interpretation of the data. In many datasets, the effect of some features are unclear to the response variable, which need to be carefully selected considering both the complexity and the accuracy of prediction model. There are many methods for feature selection, like filter method, wrapper method and embedded method [21], [22]. The embedded method incorporates the filter method and wrapper method together, which can reduce the computation cost and also consider the interaction of features.

In our scoliosis dataset, the Cobb angle is measured to reflect the severity of the spinal deformity. The other indicators reflect the states or conditions of patients. The influences of indicators on the severity of scoliosis are not clear in theory. The embedded method is utilized for the feature selection in this study. The importance of the indicators is first evaluated with Pearson correlation coefficient, and random forest importance metric then the feature selection procedures are implemented into the training procedure with 5-fold cross-validation.

1) RANDOM FOREST IMPORTANCE METRIC

Random forest [23], [24] method has been utilized to estimate the importance of features. The method evaluates the importance of a feature X_m for predicting Y by summing up the weighted impurity decreases for all nodes that used X_m, and then averaged over all trees in the forest:

$$\mathrm{Imp}(X_m) = \frac{1}{N_T}\sum_{N_T}\sum_{t \epsilon T: v(s_t)=X_m} p(t)\Delta i(s_t, t) \quad (3)$$

where $p(t)$ is the proportion N_t/N of samples reaching node t and $v(s_t)$ is the variable used in split s_t. $\Delta i(s_t, t)$ is the maximum decrease of impurity measure $i(t)$ when splitting the variable $s_t = s*$.

$$\Delta i(s_t, t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (4)$$

where the partition of theN_t node samples into t_L and t_R, and $p_L = N_{tL}/N_t$ and $p_R = N_{tR}/N_t$. The impurity measure $i(t)$ can be selected as the Gini index, the Shannon entropy or the variance of response variable Y.

2) PEARSON CORRELATION COEFFICIENT

The Pearson product-moment correlation coefficient is an index to measure the linear relation between two variables (X and Y). The Formula (5) were first developed by Pearson [25].

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (5)$$

where x_i, y_i are the realization of X and Y respectively, $\bar{x}$ is the mean of $x_i (i = 1, 2, 3, ...)$, and $\bar{y}$ is the mean of $y_i (i = 1, 2, 3, ...)$. The coefficient r can be utilized as an importance metric to measure the importance of predictor variables [21].

3) THE PROCEDURE OF FEATURE SELECTION

The embedded methods are implemented in this procedure to avoid training an exponential amount of models. The 5-fold cross-validation is utilized to train the prediction model based on different feature subsets. The performance is evaluated with mean absolute error (MAE), and root mean square error (RMSE). The procedures are shown in the following steps:

Step 1: Evaluate features' importance with random forest importance metric and Pearson correlation coefficient, then rank the features from top to bottom according to their importance respectively;

Step 2: select the most important feature as the initial subset, for example, *currentCobbAngle*, and add other features to the subset sequentially according to the ranked order;

Step 3: when one feature is added to the subset, measure the performance of the prediction model on the selected feature subset with cross-validation;

Step 4: repeat step 3 until all features are added to the subset and the performance is measured with cross-validation, compare the performance and choose the best feature subset accordingly.

E. PREDICTION WITH DATA-DRIVEN ALGORITHMS

To investigate the relationship between the recorded indicators and the future Cobb angle, both nonparametric and parametric data-driven methods are implemented for the scoliosis prediction, that are linear model (LM), k nearest neighbor algorithm (kNN), tree model, support vector machine (SVM) and random forest (RF) [26]. LM is the most basic regression method to model the linear relationship between predictor variables and responsible variable. kNN regression is a nonparametric method that predicts the value directly according to the value of nearest neighbors. The hyperparameter k is usually decided arbitrary or selected with grid search method. Tree models divide the training instances into different subspaces and make prediction with the average of the values within the same subspace. Tree models are noisy, but the bagging trees can reduce the variance greatly. The random forest is a bagging method that grows trees with random selections of the input variables, which can lessen the correlation between the trees [28]. Support vector machine (SVM) is designed for the binary classification problem by mapping features into hyperspace and then introducing the optimal hyperplane with maximized margin to separate different classes. SVM can also be adapted to regression problem by introducing the ϵ - insensitive error function for the optimization [27].

The five algorithms are the typical data mining techniques for regression, which are capable of capturing both the linear and nonlinear relationship. The prediction performance is evaluated and compared in section IV.

TABLE 4. The seven methods for dealing with missing values.

Methods	Description	Data size
A	Delete all the instances with missing values in all the selected features.	162
B	Delete one feature with most missing values, and then find the complete instances.	881
C	Delete two features with most missing values, and then find the complete instances.	1385
D	Delete three features with most missing values, and then find the complete instances.	1627
E	Keep all the selected features, impute categorical features with median values and impute numerical features with mean values respectively.	1978
F	Keep all the selected features, impute the missing values with $k$nn imputation method.	1978
G	Keep all the selected features, impute the missing values with our proposed imputation methods.	1978

F. MODEL EVALUATION

The performance of the applied models is assessed by mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and Pearson product moment correlation coefficient (r):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{9}$$

where y_i is the actual value of the Cobb angle at appointed time t_i, $\hat{y}_i$ is the prediction value of the Cobb angle at appointed time t_i, $\bar{y}$ is the mean of y_i, $\bar{\hat{y}}$ is the mean of $\hat{y}_i$, i indicates the i-th individual, n is the number of the total samples.

The MAE, RMSE, MAPE are different kinds of errors that should be minimized as much as possible, the smaller the errors, the better the model performance. While the Pearson correlation can be interpreted as a kind of accuracy, the higher the correlation, the better the model performance.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTS FOR HANDLING MISSING VALUES

In the settings of the experiments, the data are preprocessed with the above methods except for dealing with missing values. Different methods for handling the missing values are implemented for the scoliosis data, i.e. the case deletion, case deletion combined with feature deletion, median or mean value imputation and $k$nn imputation, as well as our proposed methods. It should be noted that only $k = 2$ works for the $k$nn imputation on our scoliosis data, because a larger k is unable to obtain enough complete neighbors for supporting

TABLE 5. Comparing the performance of the imputation methods.

Methods	rpMAE/rpRMSE					
	LM	kNN	TR	SVM	RF	Average
A	Selected as the baseline					
B	0.6%/-0.9%	6.1%/5.6%	13.2%/6.5%	2.4%/-0.1%	3.1%/-1.3%	5.1%/2.0%
C	0.3%/-6.7%[a]	12.7%/9.9%	15.7%/6.0%	-3.4%/-10.4%	2.0%/-7.4%	5.5%/-1.7%
D	-2.3%/-10.8%	16.8%/11%	14.0%/2.8%	0.0%/-8.7%	1.2%/-9.5%	5.9%/-3.1%
E	7.2%/-0.1%	21.8%/15.3%	19.8%/10.7%	11.4%/4.2%	14.0%/5.3%	14.8%/7.1%
F	6.6%/-0.5%	20.9%/15.3%	18.6%/8.8%	12%/4.7%	13.9%/5.2%	14.4%/6.7%
G	8.9%/3.6%	22.6%/16.6%	22.6%/11.6%	17.8%/12.1%	20.6%/12.7%	18.5%/11.3%

[a]The minus symbol means that the errors are not reduced but increased comparing to the baseline errors

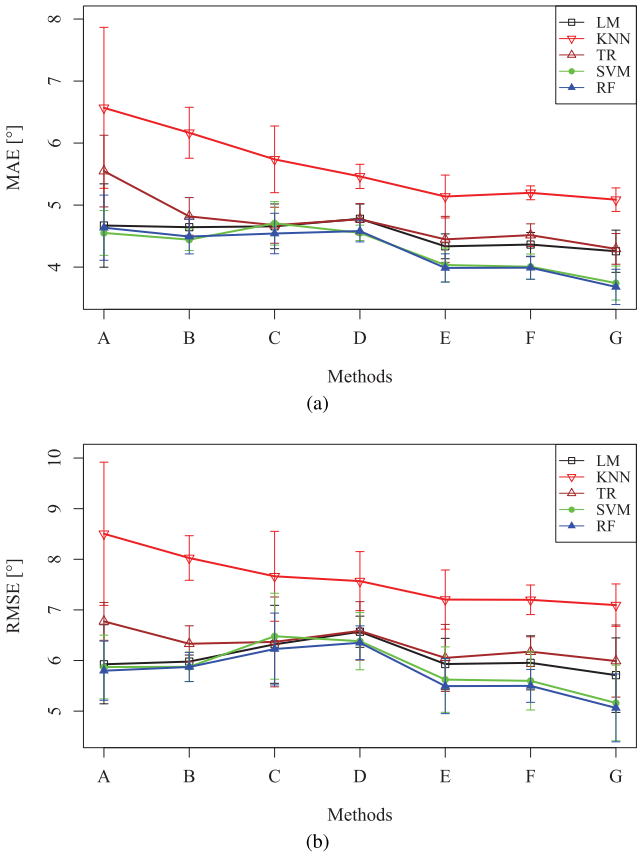


FIGURE 5. The cross-validation performance of five regression models influenced by different methods of handling missing values.

the imputation. The methods of handling missing values are summarized in details in Table 4.

The five prediction models are built on all these handled datasets. The 5-fold cross-validation MAE and RMSE are employed to evaluate the performance. The qualitative results are compared in Fig. 5. Moreover, the standard deviations among the 5-fold cross-validation are added as error bars to the mean value of MAE and RMSE. For quantitative comparing the performance of all the seven methods, the results are converted to the reduced percentage MAE or RMSE of the baseline, as shown in Formula (10) and (11). The MAE and RMSE of method A are selected as the baseline.

TABLE 6. Ranked features with decreasing importance.

Random forest	Pearson correlation
currentCobbAngle	*currentCobbAngle*
changeBrace	*currentRisserSign*
futureBrace	*currentAge*
currentRisserSign	*futureBrace*
currentBodyHeight	*changeBrace*
TimeSpan	*TimeSpan*
currentBrace	*Gender*
currentAge	*FamHx*
Gender	*currentBrace*
FamHx	*currentBodyHeight*

Table 5 shows the quantitative performances of reduced percentage MAE and reduced percentage RMSE to the baseline.

$$\text{rpMAE} = (1 - \frac{\text{MAE}}{\text{MAE}_{\text{baseline}}}) \times 100\% \quad (10)$$

$$\text{rpRMSE} = (1 - \frac{\text{RMSE}}{\text{RMSE}_{\text{baseline}}}) \times 100\% \quad (11)$$

From Fig. 5 and Table 5, our proposed approach, method G outperforms other methods in threefold: (1) the cross-validation MAE and RMSE are smaller than any of the other methods for almost all five prediction algorithms; (2) the error bars that indicate the standard deviation for our proposed methods are smaller than most of other methods for almost all the five prediction algorithms; (3) the average reduced percentage MAE (rpMAE) for all five algorithms is 18.5%, and the average reduced percentage RMSE (rpRMSE) is 11.3%, which is larger than any of the corresponding values.

B. EXPERIMENTS FOR BEST FEATURE SUBSET SELECTION

In these experiments, we evaluate the importance of features with both Pearson correlation and random forest importance metrics. Then, we utilize the embedded method to select best feature subset. The features are ranked in Table 6 according to the importance. The features at the top are considered as more important than the features on the bottom. The importance list of features are partially different due to the different criteria. Some features are in the same ranking positions while others are not, for example, the most important feature is considered as *currentCobbAngle* by both importance metric while the second important features are not the same. The two different importance rankings can provide a new perspective on the

TABLE 7. Results of feature selection.

Importance metrics	Prediction algorithms	Best feature subsets (BFS)
Random forest	LM, TR, KNN, SVM, RF	BFS 1: *currentCobbAngle*, *changeBrace*
Pearson correlation	LM, TR, KNN, SVM	BFS 2: *currentCobbAngle*, *currentRisserSign*, *currentAge*, *futureBrace*, *changeBrace*
	RF	BFS 3: *currentCobbAngle*, *currentRisserSign*, *currentAge*, *futureBrace*, *changeBrace*, *TimeSpan*

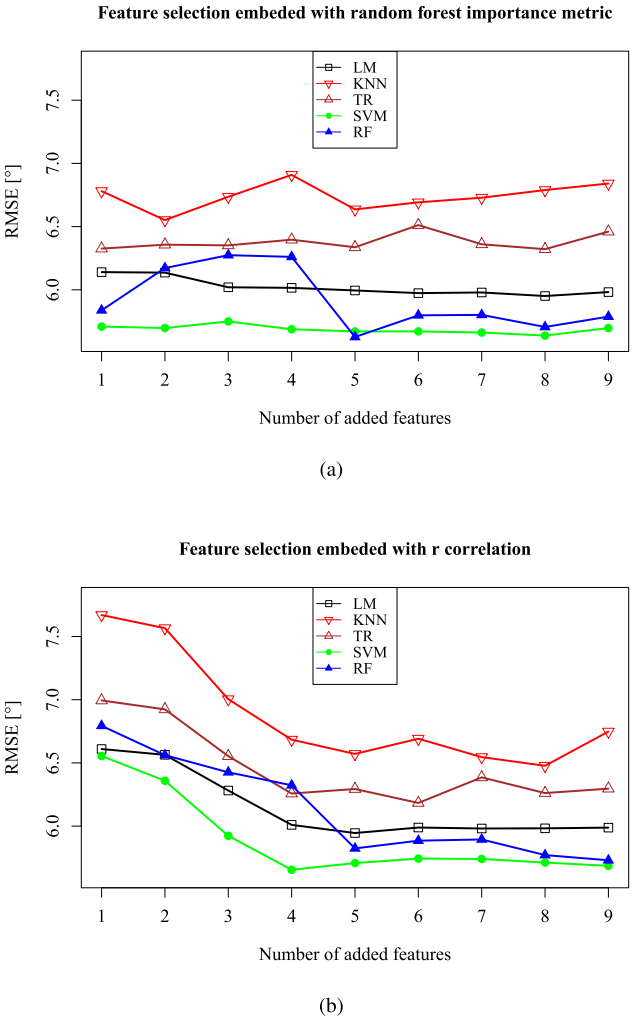


FIGURE 6. The RMSE of feature selection procedure embedded with random forest importance metrics and Pearson correlation respectively.

feature importance in terms of both linearity and nonlinearity.

The embedded method for feature selection is implemented accompanying ranked features. The performance of the five regression algorithms on all the selected feature subsets is evaluated with cross-validation RMSE and MAE. The cross-validation RMSE of the five regression models during the feature selection procedures is shown in Fig. 6. Fig. 6 (a) and Fig. 6 (b) show the RMSE corresponding to different feature importance metrics respectively. The MAE plots show the similar trend of the procedures and are not displayed here due to the limited pages.

TABLE 8. The model performance on features selected with random forest importance metric.

Features	BFS 1				
Algorithms	LM	KNN	TR	SVM	RF
MAE	4.032	4.778	4.075	3.426	3.867
RMSE	5.581	6.413	5.768	5.181	5.536
MAPE	0.246	0.267	0.256	0.216	0.243
r	0.836	0.784	0.819	0.857	0.838

From Fig. 6 (a), the RMSE is not decreased for most of the models as these features were added to the feature subset sequentially. It is hard to find the turning point that the RMSE will be significantly reduced due to the newly added feature for all five prediction models. The performance of the linear model, tree model, and support vector machine is hardly improved during the procedure; while the k nearest neighbor waves a lot during the procedure; the performance of random forest also waves at first but keeps stable at last when adding more than five features into the initial feature subset. Thus, according to the results, the first additive feature is selected to the best feature subset. The best feature subset for these five prediction models includes two features that are *currentCobbAngle* and *changeBrace*, as shown in Table 7.

The substantial decrease of RMSE for most of the models can be found in Fig. 6 (b). The RMSE is decreased for most prediction models and then keeps stable or even increase due to the overfitting as the feature adding to the feature subset. The turning point happens when adding additional four features to the initial feature subset for LM, KNN, TR and SVM. While for random forest, the turning point occurs when adding additional five features to the initial feature subset. Thus, the best feature subsets for the five models are shown in Table 7.

C. PREDICTIONS AND MODEL COMPARISONS

In this section, the latest Cobb angle is predicted for each scoliosis patient. The prediction models are built based on the selected best feature subsets. The performance is evaluated with MAE, RMSE, MAPE and Pearson correlation over the cross-validataion that leave one patient out. The results are shown in Table 8 and Table 9. The prediciton errors are shown in Fig. 7 and Fig. 8 with boxplot. In addition, the boxplot of prediction errors for best 95% samples are also indicated respectively.

From Table 8 and Table 9, the performance is similar for tree model and SVM with respect to different best feature

TABLE 9. The model performance on features selected with Pearson correlation.

Features	BFS 2				BFS 3
Algorithms	LM	KNN	TR	SVM	RF
MAE	3.844	4.55	4.074	3.41	3.632
RMSE	5.332	6.254	5.767	5.131	5.309
MAPE	0.235	0.262	0.256	0.213	0.231
r	0.848	0.791	0.819	0.86	0.852

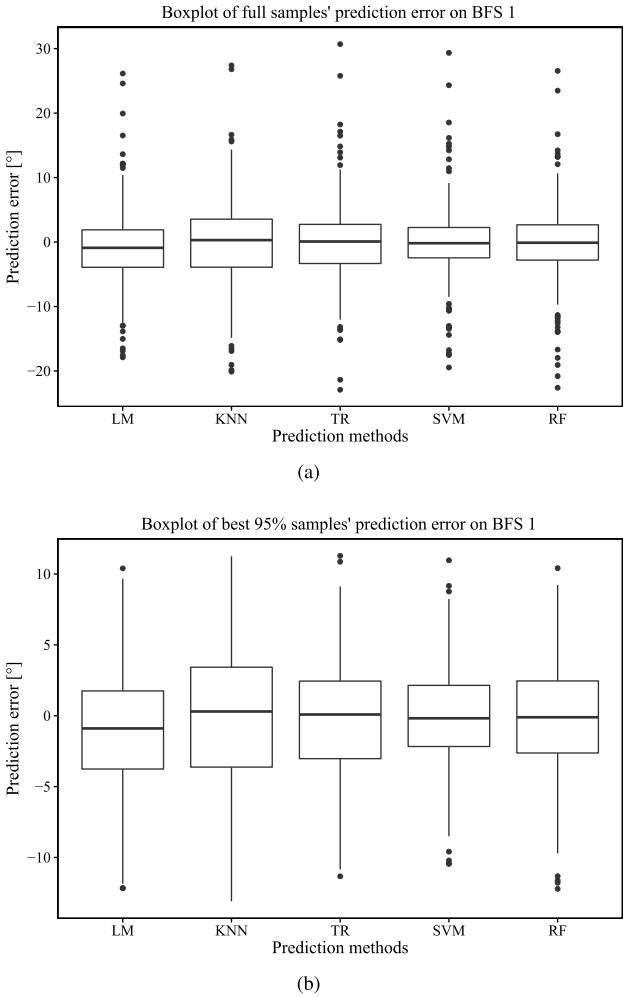


FIGURE 7. The boxplot of prediction error on the feature set selected with random forest importance metrics.

subsets. The linear model, kNN and random forest model perform slightly better on the BFS selected with Pearson correlation coefficient. However, the embedded method with random forest importance metric selects less features.

As for the individual algorithm, SVM with best MAPE 21.3% and Pearson correlation 0.86 performs slightly better than random forest and linear regression model. The linear model is quite straightforward and easy to understand. The performance of the linear model is acceptable with MAPE 23.5% and the predicted correlation 0.848. Random forest is an ensemble model, and the performance is better than the tree model as expected. kNN is not favorable for the scoliosis data due to the huge uncertainty that cannot find effective

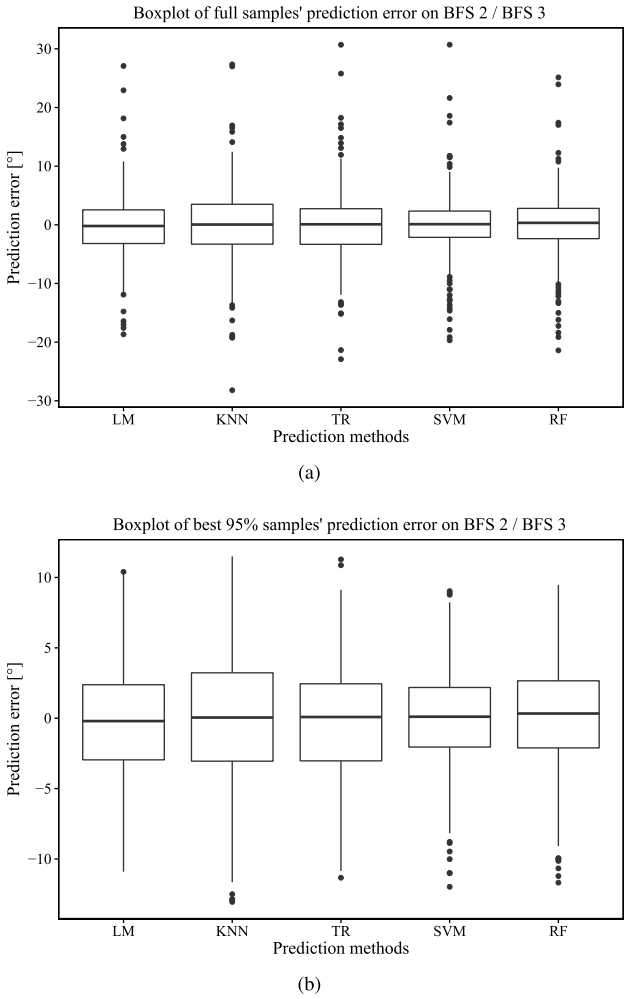


FIGURE 8. The boxplot of prediction error on the feature set selected with Pearson correlation importance metric.

nearest neighbors to support the correct predictions. SVM is suitable for an accurate prediction while linear model is good for the interpretation. The accuracy and interpretability of the models are both necessary for the decision support in the management of scoliosis.

From the boxplot of the prediction errors, the Cobb angle of some patients (less than 5%) cannot be well predicted with absolute error more than 20°. While most of the prediction errors are within ±10°, as can be found in Fig. 7(b) and Fig. 8(b).

V. DISCUSSION

Our study focuses on the prediction models to facilitate the prognosis of scoliosis in clinical workflow. The records are collected in the routine clinic setting with large uncertainty. The complexity of the scoliosis data has been analyzed and reduced with our proposed methods. The proposed missing values imputation methods try to preserve the original data information from being discarded and reduce the noise introduced by the imputed values. Different types of missing

values have been described and handled with the proposed methods. Our proposed methods outperform other deletion methods, mean/median imputation methods and *k*nn imputation method on almost all five regression algorithms.

The recorded indicators and extracted features partly reflect the initial states and the trend of these states as well as scoliosis interventions. However, the significance of these indicators with regards to progression is still unclear. In our study, we provide an importance ranking for these indicators based on an intuitive understanding of their importance with the future Cobb angle. Both linear and nonlinear importance metrics have been utilized to rank the importance of these indicators, which can provide a comparable relationship between the severity of scoliosis and these indicators. For example, we can claim that the current Cobb angle has a stronger relationship to the next Cobb angle than any other indicators in terms of both linearity and nonlinearity.

The irrelevant or redundancy information existed in the features may lead to overfitting and difficulty in interpreting the prediction models. Based on the importance ranking list, the embedded methods are employed for the feature selection to make the prediction model less complex and with higher interpretability. The performance of each prediction algorithm on feature subset that was selected with correlation coefficient is slightly better than those selected with random forest importance metric. The current Cobb angle and the change of Brace treatment are both selected by the two importance metrics and therefore are vital to the prediction of future Cobb angle. It is evident that the future magnitude of Cobb angle is highly related to the beginning magnitude of Cobb angle. The change of Brace treatment affecting the next Cobb angle is also easily understood and accepted by physicians. While the other selected features, like *currentAge*, *currentRisserSign*, and *currentBrace*, are also related to the *futureCobbAngle*, with less obvious relationship comparing the previous two features. For further investigation of these factors, the variance of the current Cobb angle and the change of Brace treatment should be carefully controlled.

Data-driven methods are promising techniques for prognosticating scoliosis. Our proposed models can effectively extract useful information from the clinical data and provide a useful tool to support the decision-making for scoliosis treatments. The heterogeneous, incomplete and random follow-up interval records can be well processed and handled with our proposed methods. With the proposed models, clinicians can compare the predicted future Cobb angles on different conditions of future brace wearing actions. Thus, physicians can make decisions based on the comparing results. The interpretability of prediction models can shed light on the prediction results and is easy to be adopted by clinicians.

VI. CONCLUSION

A decision support system is designed for scoliosis prognostication. The prediction models for scoliosis patients have been developed using various data-driven methods, which have potential to be utilized in clinical practice. The clinical data has been intensively analyzed and effectively processed to cope with the huge uncertainty. Both the current information and potential brace wearing action of patients have been extracted and incorporated into the prediction models. Features have been ranked according to the importance with respect to both linearity and nonlinearity. The embedded method for the feature selection has been implemented with importance metrics to avoid model overfitting, which also increases the interpretability of our prediction models.

The clinical data is still under collection, future work will validate the models with more clinical data and investigate the effects of other factors by controlling the initial Cobb angle and the change of Brace treatment. The prediction of double curve and triple curve are also deserving of future study.

ACKNOWLEDGMENT

The authors would like to thank Prof. Han-xiong LI for the consistent support to the project.

REFERENCES

[1] *Scoliosis in Depth Report*, accessed on Oct. 8, 2015. [Online]. Available: http://www.nytimes.com/health/guides/disease/scoliosis/print.html

[2] E. Lou, D. Hill, and J. Raso, "Brace treatment for adolescent idiopathic scoliosis," *Stud. Health Technol. Informat.*, vol. 135, p. 265, Apr. 2008.

[3] H. Wu *et al.*, "Prediction of scoliosis progression in time series using a hybrid learning technique," in *Proc. 27th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Sep. 2006, pp. 6452–6455.

[4] E. Chalmers *et al.*, "Predicting success or failure of brace treatment for adolescents with idiopathic scoliosis," *Med. Biol. Eng. Comput.*, vol. 53, no. 10, pp. 1001–1009, 2015.

[5] M. A. Asher and D. C. Burton, "Adolescent idiopathic scoliosis: Natural history and long term treatment effects," *Scoliosis*, vol. 1, no. 1, pp. 1–2, 2006.

[6] H. Wu, J. L. Ronsky, F. Cheriet, J. Harder, J. C. Küpper, and R. F. Zernicke, "Time series spinal radiographs as prognostic factors for scoliosis and progression of spinal deformities," *Eur. Spine J.*, vol. 20, no. 1, pp. 112–117, 2011.

[7] D. G. Little, K. M. Song, D. Katz, and J. A. Herring, "Relationship of peak height velocity to other maturity indicators in idiopathic scoliosis in girls*," *J. Bone Joint Surgery*, vol. 82, no. 5, p. 685, 2000.

[8] M. Ylikoski, "Height of girls with adolescent idiopathic scoliosis," *Eur. Spine J.*, vol. 12, no. 3, pp. 288–291, 2003.

[9] P. Phan, N. Mezghani, C.-E. Aubin, J. A. de Guise, and H. Labelle, "Computer algorithms and applications used to assist the evaluation and treatment of adolescent idiopathic scoliosis: A review of published articles 2000–2009," *Eur. Spine J.*, vol. 20, no. 7, pp. 1058–1068, 2011.

[10] E. Chalmers, W. Pedrycz, and E. Lou, "Predicting the outcome of brace treatment for scoliosis using conditional fuzzy clustering," in *Proc. IFSA World Congr. NAFIPS Annu. Meeting (IFSA/NAFIPS)*, 2013, pp. 837–842.

[11] P. O. Ajemba, L. Ramirez, N. G. Durdle, D. L. Hill, and V. J. Raso, "A support vectors classifier approach to predicting the risk of progression of adolescent idiopathic scoliosis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 2, pp. 276–282, Jun. 2005.

[12] L. Deng *et al.*, "Data-driven modeling for scoliosis prediction," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Sep. 2016, pp. 1–4.

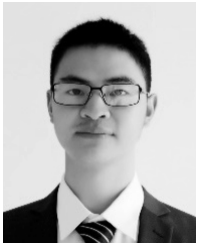
[13] L.-E. Peterson and A. L. Nachemson, "Prediction of progression of the curve in girls who have adolescent idiopathic scoliosis of moderate severity. Logistic regression analysis based on data from the brace study of the scoliosis research society," *J. Bone Joint Surgery Amer.*, vol. 77, no. 6, pp. 823–827, 1995.

[14] J. E. Lonstein and J. Carlson, "The prediction of curve progression in untreated idiopathic scoliosis during growth," *J. Bone Joint Surgery Amer.*, vol. 66, no. 7, pp. 1061–1071, 1984.

[15] Y. P. Charles, J.-P. Daures, V. de Rosa, and A. Diméglio, "Progression risk of idiopathic juvenile scoliosis during pubertal growth," *Spine*, vol. 31, no. 17, pp. 1933–1942, 2006.

[16] H.-K. Wong *et al.*, "The natural history of adolescent idiopathic scoliosis," *Indian J. Orthopaedics*, vol. 44, no. 1, pp. 1–9, 2010.

[17] S. Garavaglia and A. Sharma, "A smart guide to dummy variables: Four applications and a macro," in *Proc. Northeast SAS Users Group Conf.*, 1998, pp. 1–43.
[18] X. Feng, S. Wu, J. Srivastava, and P. Desikan, "Automatic instance selection via locality constrained sparse representation for missing value estimation," *Knowl.-Based Syst.*, vol. 85, pp. 210–223, Oct. 2015.
[19] C. Sarkar, "Improving predictive modeling in high dimensional, heterogeneous and sparse health care data," Ph.D. dissertation, Dept. Comput. Sci., Minneapolis, MN, USA, Univ. Minnesota, 2015.
[20] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.
[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.
[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[24] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 431–439.
[25] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statist.*, vol. 42, no. 1, pp. 59–66, 1988.
[26] C. M. Bishop, "Pattern recognition," *Mach. Learn.*, vol. 128, pp. 1–58, Apr. 2006.
[27] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," in *Proc. Neural Inf. Process.-Lett. Rev.*, vol. 11, no. 10, pp. 203–224, 2007.



LIMING DENG received the B.Sc. degree in mechanical engineering from Nanchang University in 2011 and the M.Sc. degree in mechanical engineering from Shanghai Jiao Tong University, China, in 2014. He is currently pursuing the Ph.D. degree with the Department of Systems Engineering and Engineering Management, City University of Hong Kong. His research interests include data mining, bioinformatics, and decision support system on clinical practice.

YONG HU (M'07–SM'11) received the B.Sc. and M.Sc. degrees in biomedical engineering from Tianjin University, Tianjin, China, in 1985 and 1988, respectively, and the Ph.D. degree from The University of Hong Kong in 1999. He is currently an Associate Professor and the Director of the Neural Engineering and Clinical Electrophysiology Laboratory, Department of Orthopaedics and Traumatology, The University of Hong Kong. His research interests include neural engineering, clinical electrophysiology, and biomedical signal measurement and processing.

JASON PUI YIN CHEUNG received the M.B.B.S. degree from The University of Hong Kong in 2007 and the MMedSc degree in 2012. He is currently a Clinical Assistant Professor with the Department of Orthopaedics and Traumatology, The University of Hong Kong. He received his training in orthopedics from Queen Mary Hospital. He completed the membership examination in 2009 and the fellowship examination in 2014. His main interests in research are pediatric growth and spinal deformity, lumbar spinal stenosis, and novel imaging.

KEITH DIP KEI LUK is currently the Chair Professor with the Department of Orthopaedics and Traumatology, The University of Hong Kong. His research interests include intraoperative spinal cord monitoring, spinal biomechanics, spinal deformity correction, genetics of scoliosis, and intervertebral disc transplantation. He is the President of the International Society for the Study of the Lumbar Spine (ISSLS), and the immediate past President of the Société Internationale de Chirurgie Orthopédique et de Traumatologie (SICOT).