



Original Article

Machine learning-based prediction of Sasang constitution types using comprehensive clinical information and identification of key features for diagnosis

Sa-Yoon Park ^{a,1}, Musun Park ^{a,1}, Won-Yung Lee ^a, Choong-Yeol Lee ^a, Ji-Hwan Kim ^b, Siwoo Lee ^{c,*}, Chang-Eop Kim ^{a,*}^a Department of Physiology, College of Korean Medicine, Gachon University, Seongnam, Republic of Korea^b Department of Sasang Constitutional Medicine, Gil Hospital of Korean Medicine, Gachon University, Incheon, Republic of Korea^c Mibyeong Research Center, Korea Institute of Oriental Medicine, Daejeon, Republic of Korea

ARTICLE INFO

Article history:

Received 15 May 2020

Received in revised form

17 September 2020

Accepted 20 September 2020

Available online 30 September 2020

Keywords:

Sasang constitutional medicine

Machine learning

Extremely randomized trees

Diagnostic model

Feature importance

ABSTRACT

Background: Despite the importance of accurate Sasang type diagnosis, a unique form of Korean medicine, there have been concerns about consistency among diagnoses. We investigate a data-driven integrative diagnostic model by applying machine learning to a multicenter clinical dataset with comprehensive features.

Methods: Extremely randomized trees (ERT), support vector machines, multinomial logistic regression, and K-nearest neighbor were applied, and performances were evaluated by cross-validation. The feature importance of the classifier was analyzed to understand which information is crucial in diagnosis.

Results: The ERT classifier showed the highest performance, with an overall f1 score of 0.60 ± 0.060 . The feature classes of body measurement, personality, general information, and cold–heat were more decisive than others in classifying Sasang types. Costal angle was the most informative feature. In pairwise classification, we found Sasang type-dependent distinctions that body measurement features played a key role in TE-SE and TE-SY datasets, while personality and cold–heat features showed importance in SE-SY dataset.

Conclusion: Current study investigated a comprehensive diagnostic model for Sasang type using machine learning and achieved better performance than previous studies. This study helps data-driven decision making in clinics by revealing key features contributing to the Sasang type diagnosis.

© 2021 Korea Institute of Oriental Medicine. Publishing services by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Sasang constitutional medicine (SCM) is a typological constitutional medicine of the Korean medical tradition that provides personalized and integrative treatment by considering both physical and psychological traits. It has been widely used in many Korean medicine clinics and hospitals and has been scientifically studied as well.^{1–3} The major feature of the SCM is that it classifies people into four Sasang types—Tae-Yang (TY), Tae-Eum (TE), So-Yang (SY), and So-Eum (SE)—based on biopsychosocial traits such as external appearance, personality, physiological symptoms, and responses

to herbal medicine treatments; thus, SCM provides a personalized treatment based on the person's Sasang type.³

Because treatment, prognosis, and health care methods are all dependent on the patient's Sasang type, accurate diagnosis of the Sasang type is a crucial step. Despite the importance of accurate Sasang type diagnosis, there have been many concerns about the consistency in the diagnosis among experts of SCM. Indeed, it was reported that the chance of three experts making the same Sasang type diagnosis is only between 52.5 and 68.4 percent.⁴ Although doctors of SCM collect and comprehensively integrate various information from the patient's whole body, a consensus has not been made on which information to emphasize more when integrating the information. For example, some experts regard body shape to be key, while others put emphasis on physiological symptoms. Thus, research on important information and reliable diagnostic criteria based on this information must be conducted.

* Corresponding authors.

E-mail addresses: bfree@kiom.re.kr (S. Lee), eopchang@gachon.ac.kr (C.-E. Kim).¹ These authors contributed equally to this work.

There have been several efforts to present quantitative diagnostic criteria for Sasang types. Do et al.⁵ attempted to develop an integrated Sasang constitution diagnosis method using face, body shape, voice, and questionnaire data, which showed 61.4% and 52.2% accuracy for men and women, respectively. Jang et al.⁶ tried to predict the Sasang type using a clinical dataset of 15 body shape features, which showed 65.1% and 63.4% accuracy for men and women, respectively. Although these studies show significantly improved performances, they also present some definitive limitations such as insufficient number of examined features, limited data compared to clinical information collected by the doctors, and arbitrary integration of different methods rather than data-driven integration.

In this study, we investigated an integrative diagnostic model using an unprecedented 248 comprehensive features, including body measurement, personality, and physiological and pathological symptom features by applying machine learning techniques. We also analyzed key features to put more weight on when diagnosing the Sasang type.

2. Methods

2.1. Data collection and preprocessing procedure

Data were obtained from the cross-sectional study conducted by the Korean Medicine Data Center (KDC) of the Korea Institute of Oriental Medicine between November 2006 and July 2013. The data includes 340 features from 3891 patients from 13 traditional Korean medicine hospitals and 11 traditional Korean medicine clinics. Diagnosis of the Sasang type was conducted by licensed medical specialists of SCM who had been in clinical practice for at least 5 years. Specialists of SCM diagnosed the Sasang type by carefully considering the physical body shape, appearance, temperament, and pathological symptoms of patients based on the detailed processes for determining Sasang types that have been previously described.⁷ All processes were approved by the Korea Institute of Oriental Medicine (I-0910/02-001) and written informed consent for participation was obtained from each subject.

To manage the missing data within the dataset, first, only the patients with data for more than 80% of the features were considered. Then, the features with data from more than 90% of the remaining patients were included in the dataset while the rest was discarded.

All twenty-five TY type patients were excluded due to their small sample size. Given the large number of features considered in our study, twenty-five samples are too small for our model to learn generalizable pattern. As a result, the final dataset of a total of 1338 patients with 248 features without missing values was used for subsequent analyses.

We analyzed the general characteristics of the 1338 subjects. The Sasang type composition of our dataset was 507 TE type (37.89%), 392 SE type (29.29%), and 439 SY type (32.81%). Our dataset was composed of 846 female (63.22%) and 492 male subjects (36.77%). The average age of the subjects was 48 (SD = 16.38; range 9–90 years). An overview of the 248 features and their classes is described in *Table S1*.

The dataset was scaled to standardize the range of individual features, except for the extremely randomized trees (ERT) classifier. In the ERT classifier, scaling was not conducted because feature scaling is not required in a tree-based model.⁸

All data preprocessing and analysis were carried out using Pandas, a Python library for data manipulation and analysis, and Scikit-learn, a Python module integrating a wide range of machine learning algorithms.⁹

2.2. Machine learning (ML) experimental details

2.2.1. ML model selection

Five well-known supervised machine learning algorithms for classification were applied for this study. The models compared in this work are as follows: ERT, linear and nonlinear support vector machine (SVM), multinomial logistic regression (Mlogit), and K-nearest neighbor (KNN).

The ERT classifier is a decision tree-based ensemble method, which is similar to random forests but uses randomly selected cut-off values rather than the optimal one. The strength of the ERT classifier is that it is robust to noise and can thus lead to a further decrease in overall variance while performing largely equal to or better than other tree-based classifiers.⁸ Furthermore, the ensemble method can be used to rank the importance of features used in a classification problem.¹⁰ The SVM classifier, one of the most powerful supervised learning techniques, searches the optimal hyperplane that maximizes the margin between classes in high-dimensional space.¹¹ The SVM classifier can be used as linear or nonlinear classifiers according to the applied kernel. For the nonlinear SVM classifier, we applied radial basis function (Gaussian kernel). To apply the SVM, a binary classifier, into a multi-class problem, one-vs-rest scheme was implied. The Mlogit classifier is used to predict a nominal dependent variable for which there are more than two categories.¹² The strength of the Mlogit model is that it not only gives a measure of how relevant a predictor (coefficient size) is, but also the predictor's direction of association (positive or negative). The KNN model is a nonparametric instance-based simple learning algorithm that classifies new cases based on a similarity measure.¹³

2.2.2. Feature selection

We calculated the adjusted mutual information (MI) to find the most relevant features for classifying the patients into a Sasang type. MI is a measure of the amount of information that one random variable contains about another random variable.^{14,15} Adjusted MI corrects for the phenomenon in which the MI is generally higher for two vectors with a larger size, regardless of whether more information is shared or not.¹⁶ For the two vectors U and V , the adjusted MI is given as:

$$\text{adjusted MI}(U, V) = \frac{[\text{MI}(U, V) - E(\text{MI}(U, V))]}{[\max(H(U), H(V) - E(\text{MI}(U, V)))]}$$

In this study, U denotes the subject vector for the Sasang type feature and V for the other features. $\text{MI}(U, V)$ is the mutual information between vectors U and V . E denotes expected value and H denotes entropy. Features that showed adjusted MI scores larger than zero (175 features) were selected to train each of the classifiers, except for the ERT classifier. We used all the features (248 features) to train the ERT classifier because decision tree-based ensemble methods can be used as feature selection methods per se.¹⁰

2.2.3. Hyperparameter optimization

For hyperparameter optimization, we conducted a grid search for hyperparameters with nested cross-validation to avoid data leakage, which is called 'nested cross-validation' or 'double cross-validation'.¹⁷ The hyperparameter values searched for each ML model are summarized in *Table S2*.

2.3. Model performance assessment

2.3.1. K-fold cross-validation

We trained each model with stratified k -fold cross-validation ($k=20$), in which the dataset is randomly divided into k disjoint

Table 1
Overall Classification Performance for Each Model

Model	Feature size	Precision	Recall	f1 score	Accuracy
ERT	248	0.600 ± 0.061	0.601 ± 0.060	0.600 ± 0.060	0.604 ± 0.060
KNN	175	0.540 ± 0.052	0.535 ± 0.054	0.538 ± 0.053	0.540 ± 0.053
Mlogit	175	0.563 ± 0.084	0.566 ± 0.086	0.565 ± 0.085	0.567 ± 0.084
SVM (linear)	175	0.531 ± 0.088	0.529 ± 0.093	0.530 ± 0.090	0.533 ± 0.088
SVM (nonlinear)	175	0.600 ± 0.054	0.600 ± 0.058	0.600 ± 0.056	0.603 ± 0.055

ERT, extremely randomized trees; KNN, K-nearest neighbor; Mlogit, multinomial logistic regression; SVM, support vector machine.

subsets of approximately equal size according to the Sasang type (Supplement 1).

2.3.2. Precision, recall, f1 score, and accuracy

We used a confusion matrix to evaluate the performance of our machine learning model, which included the precision, recall, f1 score, and accuracy.

$$\text{precision} = \frac{tp}{tp+fp}$$

$$\text{recall} = \frac{tp}{tp+fn}$$

$$\text{f1 score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

In this equation, tp denotes true positive; fp, false positive; tn, true negative; fn, false negative. Macro-average methods that treat all classes equally were used to calculate the average values in multi-class classification settings.

3. Results

3.1. Model performance for Sasang type classification

Five well-known classification ML models were applied for this study. The ML models compared in this work are the following: ERT, linear and nonlinear SVM, Mlogit, and KNN. To select the most informative features, we calculated adjusted MI and used features with adjusted MI score larger than zero (175 features). For each ML model, we report better performance between the model trained with selected features and all features. We conducted a grid search in the nested cross-validation setting for hyperparameter optimization for each model (see Methods for more details).

To evaluate the performance of each model, a 20-fold cross-validation procedure was applied to all experiments. The mean performance of the precision, recall, f1 score, and accuracy of each model is presented, with the best performance being bold faced (Table 1). The ERT and nonlinear SVM models outperformed the other models. Since the ERT model can provide more insight into important features, we decided to use the ERT model in subsequent analyses.

The classification performance of the ERT model for individual Sasang type is as follows: the average value of the precision, recall, and f1 score for the TE type were 0.74, 0.73, and 0.74, respectively; 0.76, 0.75, and 0.75 for the SE type; 0.61, 0.63, and 0.62 for the SY type (Table 2). The accuracy was 0.61 ± 0.06 (mean \pm SD), and the cross-validated macro-averaged f1 score was 0.70. The confusion matrix showed that the SY type tended to be predicted as the TE or SY types, implying that the SY type is difficult to be distinguished (Fig. 1A).

Table 2

Classification Performance of the Extremely Randomized Trees Classifier for Individual Sasang Type

	Precision	Recall	f1 score	Accuracy
TE	0.74 ± 0.05	0.73 ± 0.05	0.74 ± 0.05	0.61 ± 0.06
SE	0.76 ± 0.05	0.75 ± 0.06	0.75 ± 0.05	
SY	0.61 ± 0.06	0.63 ± 0.07	0.62 ± 0.07	
Average	0.70	0.70	0.70	

3.2. The importance of each feature in the ERT classifier

In order to understand which information plays a crucial role in Sasang type diagnosis, we analyzed the feature importance¹⁸ as determined by the ERT classifier. Fig. 1B reveals the trends of important feature classes and those of feature importance. The feature classes of body measurement (27.6%), personality (13.4%), general information (13%), and cold-heat (12.7%) were shown to play an essential role in predicting the Sasang type. Table 3 shows the features ($n=25$) that account for 25% (the 1st quartile) of the feature importance and importance values of the ERT classifier. Interestingly, the costal angle feature (2.78%) of the body measurement class was the most important feature in classifying the Sasang type (Supplement 2). Other features that had a feature importance value ≥ 1 were chest circumference, obesity, waist circumference, pelvic circumference, waist width, chest width, rib circumference. It was determined that more than half of the 1st quartile features were included in the body measurement class. Other interesting features included some personality class features, such as the speed of movement (0.81%), and the dynamic/static personality feature (0.70%), and some cold-heat class features, such as preference for cooler temperatures (0.81%) and temperature of drinking water (0.67%).

3.3. Performance variation with decremental features

To investigate how many features contributed meaningfully to the prediction of the Sasang type, we analyzed the performance while decreasing the number of features, from 248 to 1, deleting in order from the least to the most important feature. The importance of the features was calculated using an attribute of the *ExtraTreesClassifier* of Scikit-learn.⁹ From 248 to 100 remaining features, the f1 scores remained steady near 0.6; and from 100 to approximately 50 the score increased slightly to above 0.6; and after that, it dropped dramatically to 0.5 (Fig. 2). It was shown that the optimal number of features was 86 (with an f1 score of 0.611), suggesting that our dataset had extra features in classifying the Sasang type. The performance of our classifier was significantly better than the randomly permuted one over all the ranges investigated.

3.4. Binary classification between two Sasang types

Thus far, we have discussed classification of patients into the three Sasang types. However, in clinics, doctors of SCM have difficulty diagnosing between the two Sasang types after ruling out the

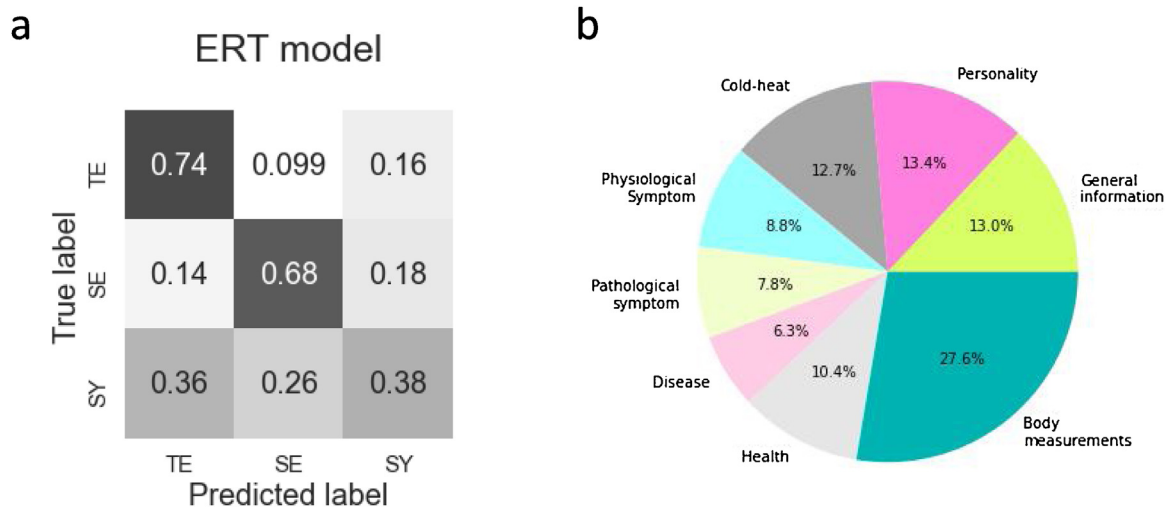


Fig. 1. Confusion matrix and feature importance of the extremely randomized trees classifier. (a) The row of confusion matrix denotes the true label in the dataset, and the column denotes the predicted label of each classifier. Each cell of the matrix indicates the average ratio value. The color of the cell indicates the value of the cell. (b) Importance value of every feature class is represented as a pie chart. Importance values for the same feature class are all summed up. TE, Tae-Eum type; SE, So-Eum type; SY, So-Yang type.

Table 3
Lists of the Features that Account for 25% of Feature Importance and Importance Values (%)

Class	Feature name	Importance (%)
Body measurement	Costal angle	2.78
Body measurement	Chest circumference	1.42
Disease	Obesity	1.39
Body measurement	Waist circumference	1.34
Body measurement	Pelvic circumference	1.22
Body measurement	Waist width	1.22
Body measurement	Chest width	1.17
Body measurement	Rib circumference	1.12
Body measurement	Axillary circumference	0.99
General information	Body weight	0.96
Disease	State of endocrinopathy	0.94
Body measurement	Hip circumference	0.94
Body measurement	Axillary width	0.82
Cold-heat	Preference for cooler temperatures	0.81
Personality	Speed of movement	0.81
Body measurement	Rib width	0.75
Body measurement	Neck circumference	0.72
Personality	Dynamic/static personality	0.70
Pathological symptom	State of digestion under bad conditions	0.68
Cold-heat	Frequency of feeling hot on your body	0.68
Physiological symptom	Temperature of drinking water	0.67
Physiological symptom	Usual degree of sweating	0.64
Personality	Active/passive personality	0.63
Physiological symptom	Degree of sweating when exercising	0.60
Body measurement	Pelvic width	0.58

least possible type in the comprehensive diagnosis process. Therefore, to identify which features should be viewed more carefully in binary classification, we also analyzed the performance and feature importance of each binary classification between TE and SE, TE and SY, and SE and SY types.

First, we analyzed each performance of these binary classifiers (TE-SE, TE-SY, and SE-SY classifier). The macro-average of the f1 score was 0.81, 0.68, and 0.73, respectively, showing the highest performance when differentiating between TE and SE types, while the lowest performance was shown when differentiating between TE and SY types (Table 4). This suggests that it is difficult to distinguish SY from TE types with our dataset. Confusion matrices were constructed for a more detailed analysis of the performance, such as true positive, false negative, false positive, and true negative rates (Fig. 3A).

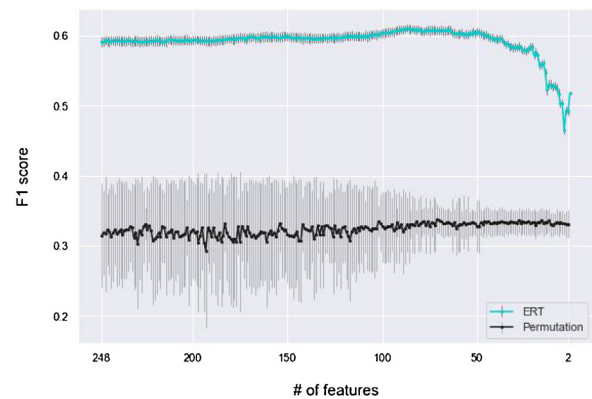


Fig. 2. Variation in performance with decremental features. Each point of upper dots (turquoise) denotes the mean of 100 f1 score when using the corresponding number of features. Each point of lower dots (black) represents the result of permutation in the same condition. Error bars represent \pm SD of the f1 scores.

In addition, to investigate which features would be decisive in Sasang type diagnosis between two types, we analyzed the feature importance in each binary classification. Table 5 shows the features that account for 25% (the 1st quartile) of the feature importance and importance values of each binary classifier. In TE-SE classification, nine features account for 25% of the feature importance, which are the costal angle, obesity, chest circumference, waist circumference, waist width, pelvic circumference, rib circumference, chest width, and state of endocrinopathy. In TE-SY classification, features with importance value ≥ 1 were the speed of movement, obesity, waist width, pelvic circumference, waist circumference, chest circumference, chest width, hip circumference, and costal angle. In SE-SY classification, features with importance value ≥ 1 were costal angle, dynamic/static personality, preference for cooler temperatures, state of digestion under bad conditions, chest circumference, and active/passive personality.

This reveals that while body measurement class features play a key role in classifying the TE-SE and the TE-SY types, personality and cold-heat class features show greater importance in classifying the SE-SY types. What is striking in TE-SE classification is that the costal angle feature alone accounts for 6.14% of the total importance value of features, and most features in the top 15 features

Table 4
Classification Performance of the Extremely Randomized Trees Classifier in Binary Classification Settings

	Precision	Recall	f1 score	Accuracy
TE	0.85 ± 0.06	0.83 ± 0.05	0.84 ± 0.04	0.81 ± 0.04
SE	0.77 ± 0.07	0.81 ± 0.06	0.78 ± 0.05	
Average	0.81	0.82	0.81	
TE	0.73 ± 0.09	0.70 ± 0.06	0.71 ± 0.07	0.685 ± 0.07
SY	0.63 ± 0.09	0.67 ± 0.08	0.65 ± 0.08	
Average	0.68	0.69	0.68	
SE	0.70 ± 0.08	0.74 ± 0.11	0.71 ± 0.08	0.73 ± 0.08
SY	0.77 ± 0.11	0.74 ± 0.07	0.75 ± 0.08	
Average	0.73	0.74	0.73	

Table 5
Lists of the Features that Account for 25% of Feature Importance and Importance Values (%) in Binary Classification Settings

TE-SE		TE-SY		SE-SY	
Feature	Importance (%)	Feature	Importance (%)	Feature	Importance (%)
Costal angle	6.16	Speed of movement	1.45	Costal angle	3.28
Obesity	2.76	Obesity	1.45	Dynamic/static personality	1.31
Chest circumference	2.53	Waist width	1.36	Preference for cooler temperatures	1.19
Waist circumference	2.34	Pelvic circumference	1.32	State of digestion under bad conditions	1.13
Waist width	2.04	Waist circumference	1.31	Chest circumference	1.10
Pelvic circumference	2.04	Chest circumference	1.26	Active/passive personality	1.06
Rib circumference	2.00	Chest width	1.17	Frequency of feeling hot on your body	0.94
Chest width	1.92	Hip circumference	1.16	Rib circumference	0.94
State of endocrinopathy	1.86	Costal angle	1.06	Temperature of drinking water	0.89
		Body weight	0.94	Extroverted/introverted personality	0.86
		Rib circumference	0.94	Pale face	0.85
		Axillary circumference	0.85	Waist circumference	0.83
		State of endocrinopathy	0.84	Self-expressive or not self-expressive	0.83
		Axillary width	0.84	Speak directly/indirectly	0.83
		Pelvic width	0.75	Bold/delicate personality	0.77
		Rib width	0.73	degree of water intake	0.75
		Usual degree of sweating	0.70	Axillary circumference	0.74
		Usual appetite	0.65	Speed of movement	0.72
		Neck circumference	0.65	Masculine/feminine personality	0.71
		Open-minded/close-minded	0.64	Pelvic circumference	0.69
		Self-expressive or not self-expressive	0.64	Decisive/indecisive personality	0.68
		Education	0.62	Thirstiness	0.67
		Coldness/warmness in hands	0.62	Chest width	0.66
		Active/passive personality	0.59	Neck circumference	0.66
		Forehead circumference	0.58	Body weight	0.66
		Decisive/indecisive personality	0.57	Feeling after sweating	0.65
		Speak directly/indirectly	0.56		
		Quantity of water intake	0.54		

are those of the body measurement class. This suggests that body shape distinguishes the TE type from SE type well, while personality class features, such as having a dynamic or static personality, are important in classifying between SE and SY types. The trends of feature importance in each classification are displayed as bar graphs in Fig. 3B.

3.5. Prediction performance comparison with a previous method

In this study, we classified the patients into one of the Sasang types using ML techniques with a dataset of unprecedented comprehensive features. We compared the performance of our method with that of the previously reported diagnostic method,⁶ which applied Mlogit model with 15 body measurement features. For the fair comparison between two methods, we constructed two different data settings using the same data source; a dataset of 1338 subjects with all the features (248 features), and a dataset of 1338 subjects with the 15 body measurement features as in the previous study. The areas under the macro-average receiver operating characteristic (ROC) curve were 0.77 for our method and 0.73 for the previous method, indicating our method is superior to the previous one (Fig. 4).

4. Discussion

The present study used machine learning techniques to investigate data-driven diagnosis criteria for the Sasang constitution type using a clinical dataset of 1338 subjects with 248 comprehensive features. According to the performance evaluation, the macro-averaged f1 score of the ERT classifier was 0.600 ± 0.060 (mean ± SD). In predicting the Sasang type, the feature classes of body measurement, personality, general information, and cold-heat played an essential role. The costal angle feature was the most important of all the features. In a binary classification setting, we found Sasang type-dependent distinctions in feature importance, where the body measurement class features played a key role in distinguishing between TE-SE and TE-SY types, while personality and cold-heat class features showed great importance in distinguishing between SE-SY types.

Several previous studies have attempted to provide quantitative diagnostic criteria for Sasang type. Do et al.⁵ attempted to design an integrated diagnostic model from four individual diagnostic models for face, body shape, voice, and questionnaire information. Although this model integrated various features, it nonetheless lacked consideration of how much each diagnostic model contributed to the integrated model, and it placed an arbitrary

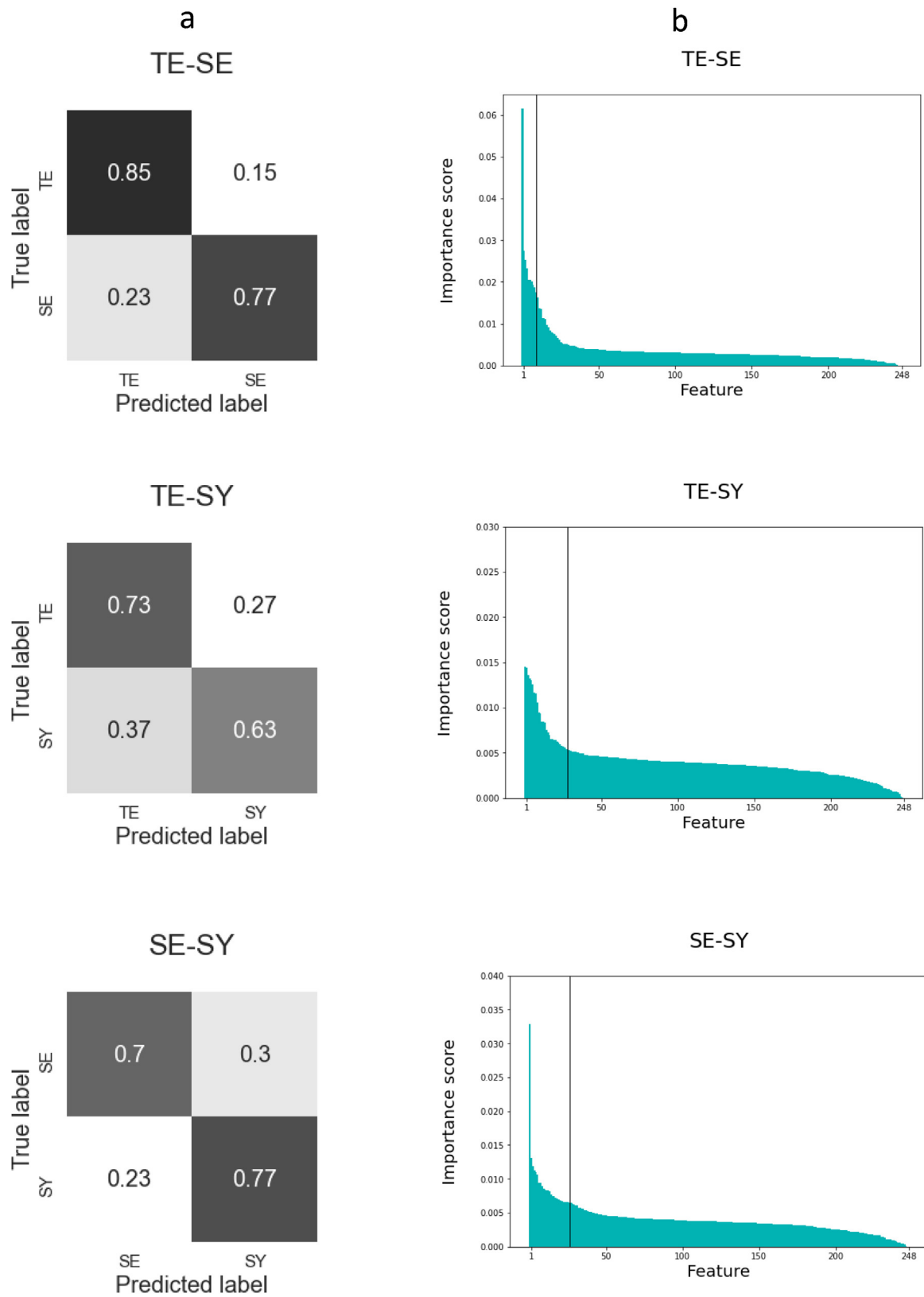


Fig. 3. Confusion matrix and feature importance of extremely randomized trees classifier in binary classification settings. (a) Confusion matrices of TE-SE, TE-SY, and SE-SY classifiers, respectively. (b) Feature importance graph in the same condition. Importance values of 248 features are sorted in descending order. The vertical line shows the 1st quartile feature. TE, Tae-Eum type; SE, So-Eum type; SY, So-Yang type.

trary equal weight on each model. On the other hand, in this study, we employed a data-driven method to reflect the importance of each feature when integrating heterogeneous multi-features. Jang et al.⁶ developed a Sasang constitutional analytic model using only

15 body measurement indices. Although this model used a data-driven method, the dataset used was limited to a small number of features, considering the kind of information doctors get in a clinical setting. Indeed, we have shown that the prediction performance

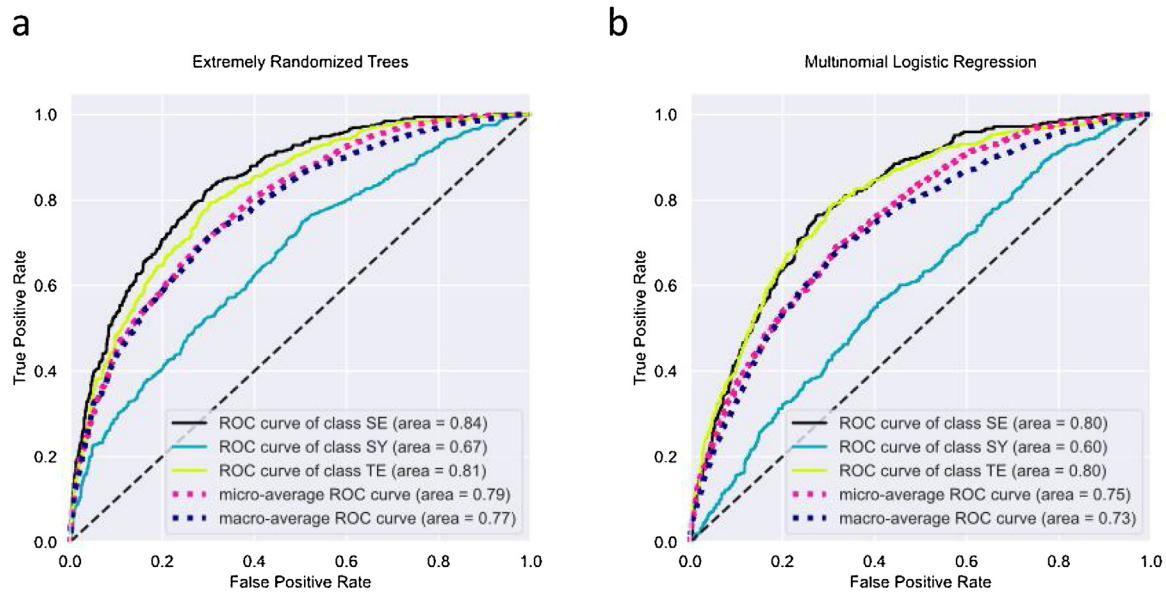


Fig. 4. Prediction performance comparison with a previous method. (a) Performance of the ERT model trained with a dataset of 1338 subjects with 248 features. (b) Performance of the Mlogit model trained with a dataset of 1338 subjects with 15 features. SE, So-Eum type; SY, So-Yang type; TE, Tae-Eum type.

of our model with comprehensive features was better than that of this model.

The present study also analyzed the feature importance of the trained model, suggesting which features to put more weight on when experts of SCM diagnose the Sasang type by integrating various clinical information. Based on the importance of each feature, the feature classes of body measurement (27.6%), personality (13.4%), general information (13.0%), and cold-heat (12.7%) turned out to play an essential role in predicting the Sasang type. Likewise, most features with high importance scores, such as costal angle, chest circumference, waist circumference, pelvic circumference, waist width, chest width, rib circumference, and axillary circumference, were all body measurement class features. This suggests the importance of body shape in diagnosing the Sasang type.

It is important to note that there are Sasang type-dependent distinctions in feature importance when discriminating between two Sasang types. When diagnosing between TE and SE, body measurement features such as costal angle, obesity, chest circumference, and waist circumference are crucial. However, when diagnosing between TE and SY types, the speed of movement feature surpassed any other body measurement feature, including the costal angle feature. When diagnosing between SE and SY, personality class features, including dynamic/static personality and active/passive personality, and cold-heat class features, such as preference for cooler temperature, were important. These differences, in turn, are evidence that there are distinctions among different constitution types as described in SCM. These results are also consistent with the explanation of Longevity and Life Preservation in Eastern Medicine,¹⁹ a text that describes physiological and pathological theories, guidelines for healthy states, and physical and psychological attributes of SCM. For example, it states that because the TE type has a small lung system and weak dispersive function, people of TE type tend to be overweight; because the SE type has a small spleen system and weak accumulating function, people of SE type tend to show poor food intake and digestion.

Although we have shown better classification performance than previous method⁶ using many more features for training classifiers, there are some limitations in this study. Due to the small sample size, the TY type patients were excluded in the analyses. Since the TY type have been considered to be rare in the population in SCM,

the problem of small sample size of the TY type is common in other studies regarding SCM diagnosis, and most of the previous studies excluded the TY type from their analyses. To improve the SC type diagnosis, it is necessary to collect more data on the TY type in future analyses.

Also, the prediction performance was limited to approximately 0.6. There are three possible explanations for this limited performance. The first explanation is the insufficiency of our data. Despite our efforts to collect as much detailed information as possible, we cannot be sure that the information in our dataset is sufficient to distinguish between the Sasang types. For instance, it may be possible that the information for patients' facial appearance, speech, or gait pattern could play a pivotal role in determining Sasang types by human experts, and these are not included in our dataset.

The second explanation is the limitation of the machine learning classifiers we used in this study. We employed various ML classifiers including SVM and decision tree-based ensemble method to predict the Sasang types, these two being among the most powerful supervised learning techniques at present. However, there remain many other classification algorithms that we have not tried here, which could yield better results than we achieved, and it may also be possible that our classifiers could be better optimized by using different data preprocessing methods or hyperparameter values.

The last and the most important possible explanation is the existence of the upper boundary of prediction performance due to the noisy and subjective inference of "ground truth"—labeled Sasang type—of the dataset by each human expert. In this study, the inference of ground truth was not cross-checked by two or more experts, although the decision was carefully made by qualified experts after evaluating Sasang type-specific drug responses. If the label feature (labeled Sasang type) between human experts is not identical, it is logical to assume that the prediction performance of the classifier is upper bounded by the agreement rate among the experts. In our previous studies,⁴ it was shown that the agreement rate for diagnosis of Sasang type was between 52.5% and 68.4% among three qualified experts. This result strongly suggests that the inferred ground truth of our dataset is also noisy and limits the prediction performance to about 60% accuracy.

The next step for objective Sasang type diagnosis should focus on improving reliability and reducing the bias in ground truth

inferences. How can patients' Sasang types be inferred in a more objective and reliable manner? Utilizing genomic or metabolomic information could be one way to achieve this goal. At the same time, it is necessary to be open to skepticism about the existence of genuine ground truth for the Sasang types. Despite the widely accepted effectiveness and the potential value of Sasang theory in personalized medicine, it is worth noting that Sasang theory was originally constructed on clinical observations of phenotypes rather than objective biological measures. It would be possible to improve and optimize Sasang theory by using objective biological data, instead of accepting the current theories as given.

Author contributions

Conceptualization: CEK, SWL and JHK. Methodology: CEK. Formal analysis: SYP and MSP. Investigation: SYP. Data curation: WYL and SYP. Writing-original draft: SYP. Writing – review & editing: SYP and CEK. Supervision: CYL and CEK. Funding acquisition: SWL and JHK.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding

This research was supported by the “Establishment of Korean Medicine Genome and Epidemiology Infrastructure” of the Korea Institution of Oriental Medicine (KIOM) Grant funded by the Korean Government (MEST) (KSN2021120), and by National Research Foundation of Korea (NRF-2017R1C1B5017048).

Ethical statement

This study was approved by the Korea Institute of Oriental Medicine (I-0910/02-001) and written informed consent for participation was obtained from each of the subjects.

Data availability

The data used to support the findings of this study were supplied by KDC of the Korea Institute of Oriental Medicine (KIOM) under license, and so cannot be made freely available. Requests for access to these data should be made at the KDC website (kdc.kiom.re.kr).

Supplementary material

Supplementary material associated with this article can be found in the online version, at [doi:10.1016/j.imr.2020.100668](https://doi.org/10.1016/j.imr.2020.100668).

References

- Kim JY, Noble D. Recent progress and prospects in Sasang constitutional medicine: a traditional type of physiome-based treatment. *Prog Biophys Mol Biol* 2014;116:76–80.
- Kim B-S, Bae HS, Lim C-Y, Kim MJ, Seo J-G, Kim JY, et al. Comparison of gut microbiota between Sasang constitutions. *Evidence-Based Complement Altern Med* 2013;2013.
- Lee SW, Jang ES, Lee J, Kim JY. Current researches on the methods of diagnosing sasang constitution: an overview. *Evid Based Complement Altern Med* 2009;6(Suppl 1):43–9.
- Baek YH, Kim HS, Lee SW, Jang ES. The concordance and validity assessment of diagnosis for the expert in Sasang constitution. *J Sasang Const Med* 2014;26:295–303.
- Do JH, Jang E, Ku B, Jang JS, Kim H, Kim JY. Development of an integrated Sasang constitution diagnosis method using face, body shape, voice, and questionnaire information. *BMC Complement Altern Med* 2012;12:85.
- Jang E, Do JH, Jin H, Park K, Ku B, Lee S, et al. Predicting sasang constitution using body-shape information. *Evid Based Complement Altern Med* 2012;2012:398759.
- Song KH, Yu SG, Cha S, Kim JY. Association of the apolipoprotein A5 gene -1131T>C polymorphism with serum lipids in Korean subjects: impact of Sasang constitution. *Evid Based Complement Altern Med* 2012;2012:598394.
- Geurts P, Ernst D, Wehenkel LJ. Extremely randomized trees. *Mach Learn* 2006;63:3–42.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer; 2008:313–25.
- Steinwart I, Christmann A. *Support vector machines*. Springer Science & Business Media; 2008.
- Böhning D. Multinomial logistic regression algorithm. *Ann Inst Stat Math* 1992;44:197–200.
- Fukunaga K, Narendra PM. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Trans Comput* 1975;100:750–3.
- Liu H, Liu L, Zhang H. Feature selection using mutual information: an experimental study. In: *Pacific rim international conference on artificial intelligence*. Springer; 2008:235–46.
- Zhang Z, Hancock ER. Mutual information criteria for feature selection. In: *International workshop on similarity-based pattern recognition*. Springer; 2011:235–49.
- Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J Mach Learn Res* 2010;11:2837–54.
- Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform* 2014;6:1–15.
- Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform* 2009;10:213.
- Lee J-M. *Longevity and life preservation in Eastern medicine*; 1894.