Doctoral Dissertation

Cross-Algorithm Validation for Predicting Body Constitutions by Life-Style Towards Bias Recovery

Guang SHI

Program of Information Science and Engineering
Computational Systems Biologoy Lab. (Division of Information Science)

Supervisor: Professor Shigehiko Kanaya
Graduate School of Science and Technology
Nara Institute of Science and Technology

September 14, 2021

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Guang SHI

Thesis Committee:

Supervisor Shigehiko Kanaya
(Professor, Division of Information Science)
Keiichi Yasumoto
(Professor, Division of Information Science)
MD.Altaf-Ul-Amin
(Associate Professor, Division of Information Science)
Naoaki Ono
(Associate Professor, Division of Information Science)
Alex Ming Huang
(Assistant Professor, Division of Information Science)

Cross-Algorithm Validation for Predicting Body Constitutions by Life-Style Towards Bias Recovery*

Guang SHI

Abstract

The body constitutions (BCs) analysis is one of critical issues in the traditional Chinese medical (TCM) theory for the diagnosis, therapy, and healthcare guidance. The categorization of BCs based on symptoms, known as BC diagnosis, has been well investigated through clinical questionnaires. However, the prediction of BCs by the daily life-styles is still demanded. By crowdsourcing of effective life-style questionnaires, the machine learning (ML) algorithms are expected to precisely predict the BCs without symptom investigations. Beyond the accuracy of ML prediction, it is necessary to effectively extract critical life-style items from a large amount in the questionnaire since the clinical doctors can only suggest a limited number of life-style for adjustment, practically.

Methods: A well-agreed Japanese Version of Constitution in Chinese Medicine Questionnaire (CCMQ-J) is assigned to 851 persons for BC diagnosis as a refereed criterion. Another questionnaire including 254 life-style items is created and distributed to the same individuals as above. Harvesting the answers through crowdsourcing, multiple machine learning algorithms are applied to predict the BC of each individual on the basis of life-styles. The algorithms including random forest (RF), partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), and a novel pair-wise scheme of LASSO are adapted for cross-validation. Moreover, the principle features of life-style for each BC are extracted and analyzed by all the algorithms.

*Doctoral Dissertation, Graduate School of Information Science,
Nara Institute of Science and Technology, September 14, 2021.

Results: By predicting the BCs through RF algorithm, the prediction accuracy of 88.7% is achieved with the full items of life-style questionnaire. The applications of PLS and LASSO regression perform the prediction accuracy of 40.9% and 69.9%, respectively. However, the principle features of life-style items are extracted into an average amount of 28 and 31, respectively. The proposed pair-wise scheme of LASSO improves the prediction accuracy to 94.6% with 17 principle features in average. All the principle features obtained by the employed algorithms are summarized, compared and analyzed in terms of TCM theory.

Conclusion: The life-style is feasible to predict BCs by the symptom-blind validation. Our designed questionnaire for life-style is effective. Furthermore, the proper use of ML algorithms leads to high-quality prediction with greatly reduced number of principle life-styles, which helps to recover biased BCs into a gentle constitution by adjusting reasonably few life-styles. Most of extracted principle features by scientific data-analysis are explainable and addressable to the TCM theory.

Keywords:

body constitution, traditional Chinese medicine, life-style, bias recovery, machine learning prediction

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

In many fields of complementary and alternative medicine, the body constitution (BC) is considered as the basis of therapy and health care [6, 65]. A systematic theory of BC categorization and its application have been well investigated in the traditional Chinese medicine (TCM) for centuries [69], especially. In such a theory, the body constitutions (BCs) of individuals are categorized into nine types [72], which include one balanced type of gentleness (GN), and other eight biased types, (i) Qi-deficiency, (ii) Yang-deficiency (YA), (iii) Yin-deficiency (YI), (iv) Phlegm-wetness (PW), (v) Wet heat (WH), (vi) Blood-stasis (BS), (vii) Qi-depression (QD), and (viii) Special diathesis (SD). This fashion of BC categorization has been proven effective for clinical diagnosis, treatment, and therapy [27, 30, 35, 76]. On the other hand, identifying individuals into proper BCs is hardly conducted by observing some specific physical or chemical indexes straightly. In the TCM theory, the BCs are diagnosed by analyzing various symptoms of individuals through the clinical inquiry [7, 60, 71], which is also seen as the “questionnaire" in ubiquitous data science. It has been widely agreed the series of constitution of Chinese medicine questionnaire (CCMQ) along with its scientific categorization rule is a standard criteria for BC diagnosis [70, 88, 90]. The CCMQ has been adapted in various regions and nations with necessary adjustment [43, 44, 75]. For instance, the Japanese version of CCMQ (known as CCMQ-J) with 60 items was developed and proved effective in Japan [78, 89]. By using the CCMQ series, the body constitution is conveniently diagnosed according to the observed symptoms (addressing the items in the questionnaire) through the specific scoring rule, which is calculated by a set of linear equations from TCM theory. On the other hand, it is also important to “predict" the body

constitutions by the life-style before the symptoms appear.

In our early investigation [59], the BC prediction was conducted by a novel life-style questionnaire with 254 items. This questionnaire covers seven aspects of human daily life as (i) psychological distress [15, 32], (ii) job stress [25, 54], (iii) the Big Five personality [26, 53], (iv) sleep quality [13], (v) living conditions [1], (vi) nutritional appetite [52], and frequency of food selection [81]. Through crowdsourcing [5, 20], the life-style questionnaire along with CCMQ-J is distributed to the same objective population, where the CCMQ-J diagnosis results are only referred as the BC criteria. Since the number of items is much larger than CCMQ, the linear calculation is infeasible to the prediction by life-styles. The advanced mathematical approaches such as machine learning are necessary. It has been found that BCs can be predicted by the life-style through some specific machine learning algorithms. However, the prediction accuracy is expected to improve.

From the clinical point of view, the simple BC prediction by life-style questionnaire is meaningless since the BC of individual can be conveniently diagnosed by much simpler symptom questionnaire such as CCMQ. Beyond the prediction accuracy, a great challenge lies on guiding individuals with biased BCs to recover into the gentleness constitution practically. Therefore, the reasonable set of principle features (PFs) among massive life-style items are expected to identify for distinguishing the biased BCs and the gentleness. The quality, amount, and reliability of PF identification should be considered in clinical field. Any one of specific ML algorithms hardly offers convincing and reliable solution. A multiple algorithm cross-validation is demanded in the sense of implicit big-data analysis.

1.2 Machine Learning Expert System in Clinic Field

The ariticial intelligence (AI) has been widely developed and applied in various fields in past decades [4, 12, 28, 46, 57, 64, 77]. As one of most important aspects of AI, the machine learning technology is considered as the powerful tool to analyze big-data and assist human to solve complex problems in real-world [29, 38, 83]. For instance, many remarkable works have been reported to apply

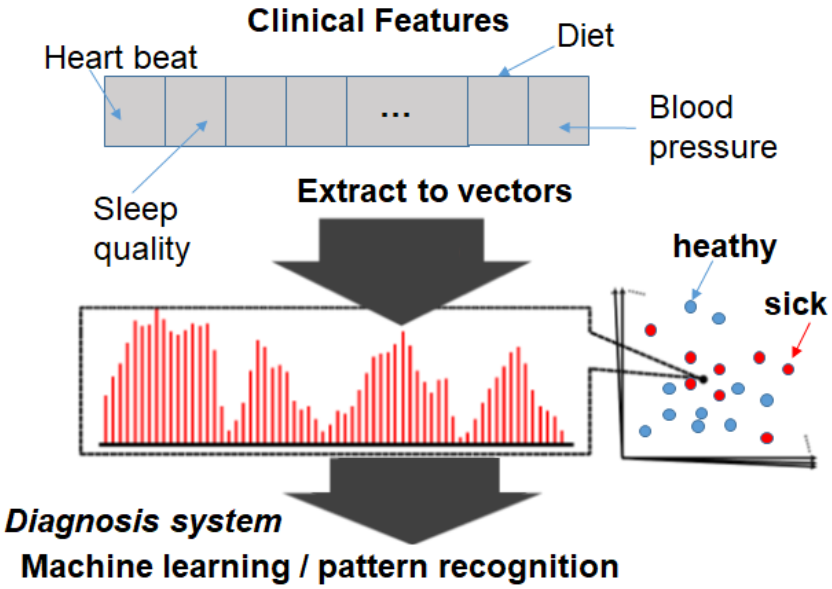

Figure 1.1: Physical indexes such as symptoms are extracted into vectors and fed into a machine learning classification/regression system for diagnosis or prediction.

machine learning technologies in the medical science such as disease diagnosis systems [9, 16, 33, 36, 55, 84, 87]. Since most of diagnosis systems are built through data classification or so-called pattern recognition networks, the expert system is expected to apply some recently developed machine learning algorithms in general purpose. From some laboratory works in medical science, the classification algorithm performs very high correct rate in categorizing medical cases which is described by the high-dimensional feature vectors [2, 66, 73]. Employing a sufficiently large database of disease cases with correct labels, those classifiers are built by training process. Then, the well-trained classifier accepts new cases of which the category labels are unknown ("ill" or "healthy" for instances). The classifier, known as diagnosis system, gives the predicted labels instead of clinical doctor' judge as shown in Fig. 1.1.

However, the reality is that no any clinical doctor makes critical judgement through AI instead of his/her expertise knowledge. Namely, what we really need is NOT a diagnosis system but a diagnosis assistant

system as shown in Fig. 1.2. In this thesis, a BC prediction system is proposed for BC status categorization for assisting the clinic doctors to make life-guidance. Namely, the purpose is to extract some principle information by the machine learning system instead of making the system communicate with the individuals/patients. For most of medical expert systems, the big data is

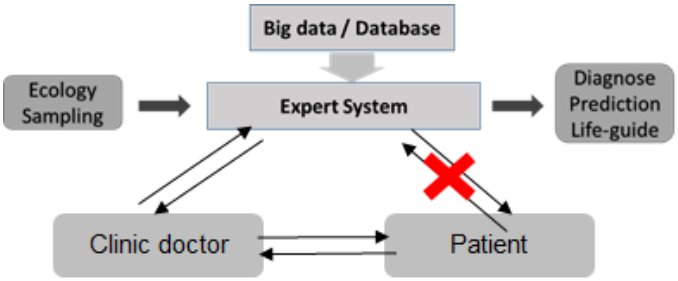

Figure 1.2: Expert system for medical analysis is supported by big data and assisting clinic doctors for diagnosis and prediction.

one of most critical issues on the quality of entire systems. In general, both amounts of sample size and dimensions are massive (or expected to be massive). An important task of ML expert system is to compress those amounts into a reasonable scale with highly effective information. From the engineering point of view, this task is seen as sample compression and dimension reduction. Our implementation in this thesis is actually one trial of dimension reduction. From the massive life-styles, the most effective items are expected to identify. The great challenge lies on that our sample data is quite limited but our expected amount on "effective items" are quite small.

Therefore, no any specific ML algorithm is feasible to achieve our goal mentioned above. Among hundreds of items, only several "effective items" are expected to perform by using hundreds of samples. A special strategy is proposed by using multiple ML algorithms with differing features for cross-validation; then, the common indications from them are summarized towards our goal.

1.3 Traditional Chinese Medicine and body consitution

Traditional Chinese Medicine (TCM) has been developed and applied for dozen of centuries. It was mainly proven and studied by Chinese (or eastern Asian) people [40, 41, 45, 47, 48, 49, 50, 61, 62, 85]. Limited by the poor information technology of human being in such a long history, there are not rich datum in the sense of scientific fashion remaining on TCM. Thus, the TCM appears quite literal and experience-driving scope [3, 8, 11, 19, 31, 37, 42, 58, 63, 79, 80, 82, 86]. A remarkable part of definition, derivation, data, and experiments results are given in terms of literal comments or implicit indications. It leads to the difficulty on the scientific data analysis on the basis of those theories. The machine learning algorithms, which are capable to process implicit datum, are helpful to verify, investigate, even develop TCM theory due to their special property [14, 21, 22, 23, 39, 74]. As mentioned as above sections, the body constitutions are important

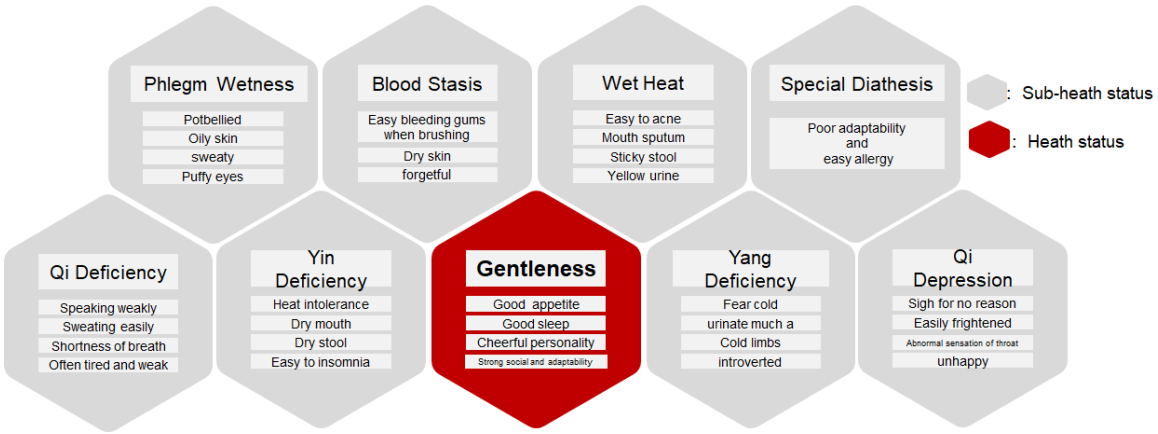

Figure 1.3: BC diagnosis and description are given in literal comments briefly in original TCM theory.

field of TCM, which typically appear the implicit characteristics of TCM. Figure 1.3 describes this characteristics in details. It is seen that the diagnosis and description of each BC are given very literally and briefly in the original TCM theory.

1.4 Overview of this work

This work is a research project to offer the life guidance for balancing the body constitution, which is a cross-field effort between medical and data science. As a candidate of complementary medical fields, traditional Chinese medical theory appears scientific and gentle effects on the diagnosis, prediction, and healthcare. In such a theory, the body constitutions (BCs) of individuals are widely considered as the most important indexes for diagnosis and treatment. This research aims at precisely categorizing individuals into specific BCs out of nine types from the evidence of physical indexes; predicting the BCs from the life-style; offering the health guidance on the life-styles for recovering the so-called "biased" BCs to the healthy status known as the "Gentle BC". By using the real-world data-set from various questionnaires, the key features of life-style are identified through machine learning (ML). However, the conventional ML algorithm for such application, known as random forest (RF), hardly offers a small set of significant life-style features. In this sense, this effort of data-analysis has almost no practical impact to clinical fields. In this project, the investigator analyzed the clinical data through various ML algorithms including RF, partial least squares (PLS) regression, least absolute shrinkage and selection operator (LASSO), even developed a novel scheme of LASSO; then, refined the medical evidences from all the ML results. Finally, the general life-style guidance is given.

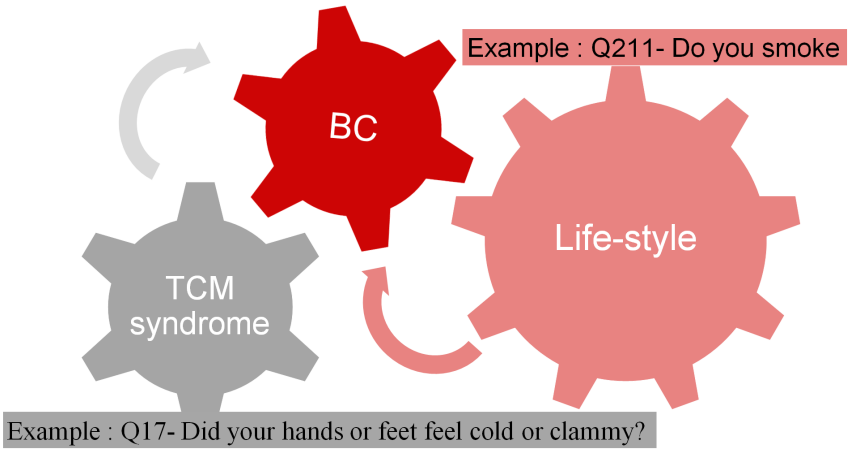

Figure 1.4: Overview of this work.

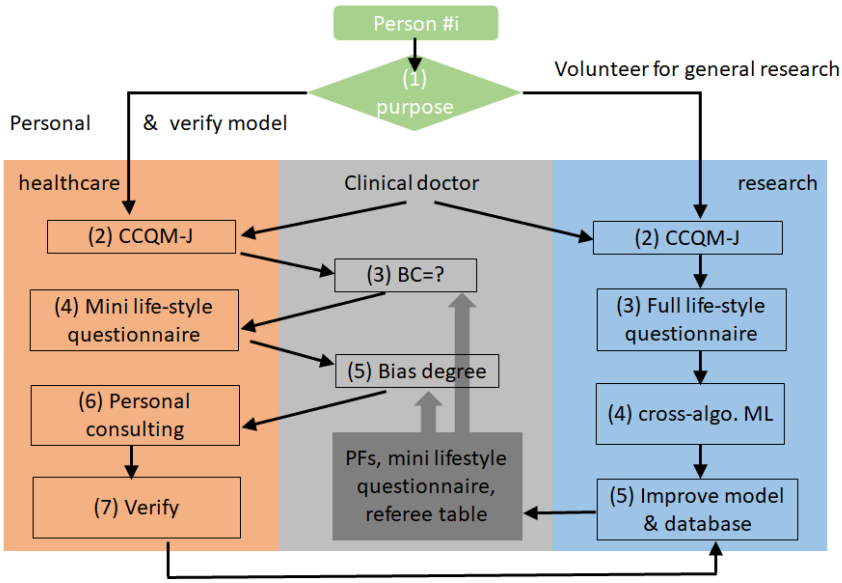

Figure 1.5: Implementation flow of proposed BC-treatment system

The implement flow of proposed BC-treatment system is illustrated as Fig. 1.5. The principle of such a system lies on distinguishing the application purpose of individuals. Here, two types of scopes are clarified: one is for the personal consulting and bias-recovery and another one is the volunteers for improving the data-base and ML models on the research purpose. For the former one, the clinical doctor is expected to refer the relevant models (principle features, bias degree etc., seeing the details of rest parts of this thesis) and offer the personal consulting according to CCMQ-J answers and a mini life-style questionnaire. For the latter one, the volunteers are expected to feed their full answers of both questionnaire into the data-base and rebuild the ML model incrementally.

1.5 Organization of this Thesis

The rest part of this thesis is organized as follows: Chapt. 2 briefly introduces the diagnosis flow of BC diagnosis by CCMQ-J as the preliminary of this work. The machine learning based BC prediction through life-styles is proposed in Chapt. 3, where the life-style questionnaire and multiply algorithms such as RF, PLS, LASSO, and a new scheme of pair-wise classification are explained in detail. In

Chapt. 4, the prediction results and principle features of all the algorithms are illustrated. Chapt. 5.3 offers the cross-algorithm validation and health-guidance for recovering the biased BCs into GN. The Conclusions are made in Chapt. 6. Following the acknowledgement in Chapt. 7, appendix A shows the abbreviations in this thesis; and appendix B introduces a machine learning method to identify the principle samples, which will be meaningful when the BC-relevant data increases greatly.

2 Related Work

The diagnosis of BC is seen as the parent work of this thesis. According to the shape of the human physical, functional, psychological, and other characteristics, an individual constitution can be assessed by the Constitution in Chinese Medicine Questionnaire (CCMQ) developed by Wang et al [70, 88, 90] which has been translated in Japanese called Japanese version of CCMQ (CCMQ-J) [78, 89]. The nine constitution types were classified into a balanced constitution (i.e. Gentleness type, abbreviated as GN), and the eight types of unbalanced types, (i) Qi-deficiency, (ii) Yang-deficiency (YA), (iii) Yin-deficiency (YI), (iv) Phlegm-wetness (PW), (v) Wet-heat (WH), (vi) Blood-stasis (BS), (vii) Qi-depression (QD), and (viii) Special diathesis (SD).

As the preliminary of this work, the BC should be diagnosed in the scientific term. Our prediction performance will be verified by those diagnosis as the "teacher" labels. In this sense, the BC diagnosis is expected to be offered in the highly trustable manner on the side of medical science instead of data science. Figure 2.1 illustrates the flow of diagnosis.

2.1 Briefing CCMQ-J

The nine types of body constitutions of CCMQ-J are derived from Likert scales of 60 questions from 1 to 5 assigned as never, rarely, sometimes, often and always, respectively. Scores of the nine constitutions are calculated by nine equations for the body constitutions using the scores of the 60 questions.

Currently, Meikirch proposed the definition of health integrating a new view in 2014 summarized as follows [6], "Health is a dynamic state of well-being emergent from conductive interactions between an individual's potentials, life's demands, and social and environmental determinants. Life's demands can be physiolog-

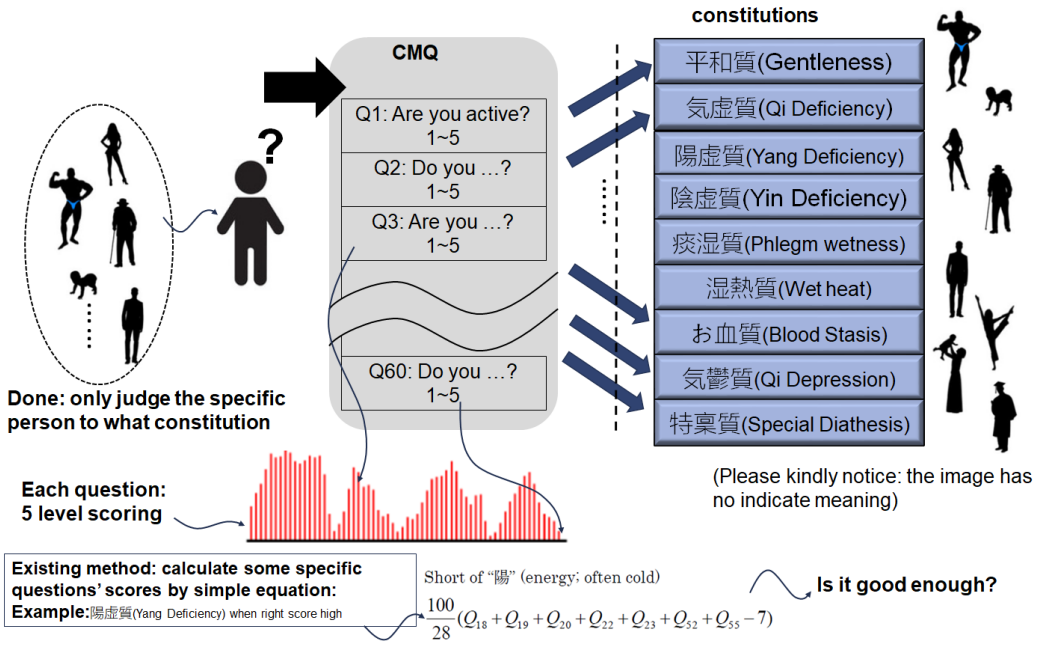

Figure 2.1: Diagnosis flow by quantitizing answers of individuals into computable data

ical, psychosocial, or environmental, and vary across individuals and contexts, but in every case, unsatisfactory responses lead to disease." The theory of TCM constitution provides personalized services and requires different treatment, for example, different food due to different food properties, different guidance on work and rest and on exercise regimen, and different Chinese herbs 9). By combining the concepts of CCMQ and Meikirch model, it should be expected that constitution scores determined by CCMQ-J can be explained by multifaceted factors, i.e., physiological, psychosocial, environmental factors and diet habits. So we try to explain the scores of the nine types of constitutions in CCMQ-J by 851 participants recruited by crowdsourcing based on several questionnaires concerning health balance, i.e., (i) psychological distress, (ii) job stress, (iii) the Big Five personality, (iv) sleep quality, (v) living conditions, (vi) nutritional appetite and frequency of food selection.

The entire quetionnaire is given as Tab. 2.1

Table 2.1: Full version of CCMQ-J

ID	question contents
C1	SH: Were you energetic?
C2	SH: Did you get tired easily?
C3	SH: Did you suffer from shortness of breath?
C4	SH: Do you get nervous easily?
C5	SH: Did you get dizziness easily or become giddy when standing up?
C6	ME: Did you prefer quietness and do not like talk?
C7	SH: Do you feel feeble when talking?
C8	ME: Did you feel gloomy and depressed?
C9	ME: Do you get nervous and worried easily?
C10	ME: Do you feel sensitive, vulnerable or emotionally upset?
C11	ME: you easily scared or frightened?
C12	SH: Did you experience distention in the underarm or breast?
C13	SH: Did you feel chest or stomach stuffiness?
C14	ME: Did you sigh for no reason?
C15	SH: Did your body feel heavy or lethargic?
C16	SH: Do the palms of your hands or soles of your feet feel hot
C17	SH: Did your hands or feet feel cold or clammy?
C18	SH: Did you feel cold easily in your abdomen, bac k, lower back or knees?
C19	SH: Were you sensitive to cold and tend to wear more clothes than others?
C20	SH: Did your body and face feel hot?
C21	SH: Did you feel more vulnerable to the cold than others
C22	SH: Did you catch colds more easily than others?
C23	SH: Did you sneeze even when you did not have a cold?
C24	SH: Did you have runny or stuffy nose even when you did not have a cold?

Continued on next page

Table 2.1 – continued from previous page

ID	question contents
C25	SH: Did you cough due to seasonal change, temperature change, or unpleasant odor?
C26	SH: Did you sweat easily when you had a slightly increased physical activity?
C27	SH: Did you forget things easily?
C28	SH: Did you have an excessively oily forehead and/or t-zone?
C29	SH: Were your lips redder than others?
C30	SH: Do you get allergies easily(are you allergic to medicine, food, odor, flower powder, or during seasonal or weather change)?
C31	DI: Did your skin get hives/ urticaria easily?
C32	DI: Did your skin have purpura (purple spots, ecchymosis) due to allergies?
C33	SH: Does black or purple ecchymosis(spots) appear on your skin for no reason?
C34	SH: Did your skin turn red and show traces when you scratched it?
C35	SH: Did your skin or lips feel dry?
C36	SH: Did you have visible capillary thread veins on your cheeks?
C37	SH: Did you feel pain somewhere in your body?
C38	SH: Do you have blush or red traces on your cheeks(hot flashes)?
C39	SH: Did your nose or your face feel greasy, oily, or shiny?
C40	SH: Did you have a dark face or get brown spots easily?
C41	SH: Did you get acne or sores easily?
C42	SH: Did you have upper eyelid swelling?
C43	SH: Did you get dark circles under the eyes easily?
C44	SH: Did your eyes feel dry and use eye drops?
C45	SH: Did your lips darker, more blue or purple than usual?
C46	SH: Did you often feel parched and need to drink water?
C47	SH: Did your throat feel strange(i. e., Like something was stuck or there was a lump in your throat)?
C48	SH: Did you have bitterness or a strange taste in your mouth?

Continued on next page

Table 2.1 – continued from previous page

ID	question contents
C49	SH: Did your mouth feel sticky?
C50	SH: Was your stomach/belly flabby?
C51	SH: Did you have lots of phlegm, especially in your throat?
C52	SH: Did you feel uncomfortable when you drank or eat something cold, or do you avoid to drinking or eating something cold?
C53	SO: Could you adapt yourself to external natural or social environment change?
C54	SL: Did you suffer from insomnia?
C55	SH: Did you easily contract diarrhea when you were exposed to cold or eat(or drink) something cold?
C56	SH: Did you pass sticky stools and/or feel than your bowel movement is incomplete?
C57	SH: Did you get constipated easily or have dry stools?
C58	SH: Did your tongue have a thick coating?
C59	SH: Did your urethral canal feel hot when you urinated, or did your urine have a dark color?
60a	SH: Was your scrotum always wet(only for male interviewees)?
60b	SH: Was your vaginal discharge yellowish(only for female interviewees)?

2.2 BC Diagnosis through CCMQ-J

Harvesting all the answers from 851 persons through crowdsourcing, the five-level indexes are calculated in the linear functions on the basis of TCM theory. Namely, the analysis of CCMQ-J only refers the medical science instead of data science. Corresponding to each body constitution among nine types, nine scores are given by above calculations. The calculation rule is illustrated in Tab. 2.2. After scoring 851 persons' BC evaluation, the winner-take-all mechanism is applied to judge the proper BC categorization as shown in Fig. 2.2. Among nine scores of

a specific person, the highest score is selected to diagnose his/her BC type.

Table 2.2: Equations for calculating BC scores for each type

Body constitution	equation for scoring
GN	$\frac{100}{32}(Q_1 + Q_{53} - Q_2 - Q_7 - Q_8 - Q_9 - Q_{22} - Q_{54} - 28)$
QF	$\frac{100}{32}(Q_2 + Q_3 + Q_4 + Q_5 + Q_6 + Q_7 + Q_{23} + Q_{27} - 8)$
YA	$\frac{100}{28}(Q_{18} + Q_{19} + Q_{20} + Q_{22} + Q_{23} + Q_{52} + Q_{55} - 7)$
YI	$\frac{100}{32}(Q_{17} + Q_{21} + Q_{29} + Q_{35} + Q_{38} + Q_{44} + Q_{46} + Q_{57} - 8)$
PW	$\frac{100}{32}(Q_{14} + Q_{16} + Q_{28} + Q_{42} + Q_{49} + Q_{50} + Q_{51} + Q_{58} - 8)$
WH	$\frac{100}{24}(Q_{39} + Q_{41} + Q_{48} + Q_{56} + Q_{59} + Q_{60} - 6)$
BS	$\frac{100}{28}(Q_8 + Q_{33} + Q_{36} + Q_{37} + Q_{40} + Q_{43} + Q_{45} - 7)$
QD	$\frac{100}{28}(Q_9 + Q_{10} + Q_{11} + Q_{12} + Q_{13} + Q_{15} + Q_{47} - 7)$
SD	$\frac{100}{28}(Q_{24} + Q_{25} + Q_{26} + Q_{30} + Q_{31} + Q_{32} + Q_{34} - 7)$

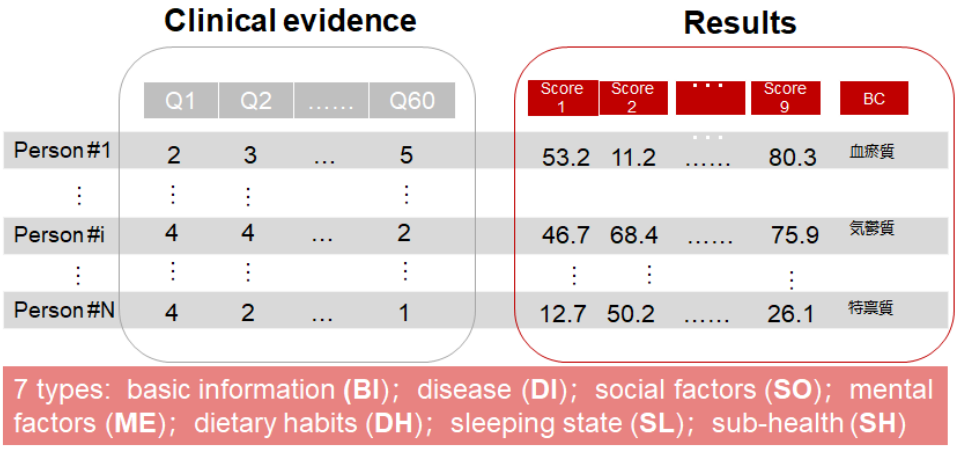

Figure 2.2: Scores from CCMQ-J calcluations are employed to perform diagnosis through winner-take-all mechanism.

In fact, the diagnosis of BC is out of scope of this study. However, the results of diagnosis is referred as the teachers' sample in the post process, prediction for instance. It is found that not only the category labels but also the scores are offered by the CCMQ-J. Therefore, it is convenient to choose the prediction algorithms somehow. Both of regression and classification can be applied to categorize. In the rest part of this thesis, those two strategies are employed.

3 Proposed Methods for Predicting BCs by Life-Style

3.1 Framework

This work aims at predicting the body constitutions by life-style instead of observing symptoms; and identifying the key factors of massive life styles for recovering the biased BCs into the gentleness. The overview is seen in Fig. 3.1. The categorization of body constitutions is usually conducted by the diagnosis of symptom inquiry in clinical field. A questionnaire such as CCMQ is feasible to perform the inquiry on the basis of easily understandable questions to ordinary individuals. In this work, the CCMQ-J [78, 89] is adapted and distributed to 851 answers (350 males and 501 females between 20 and 85 years old) via crowdsourcing. In such a diagnosis scheme, each question is quantitatively evaluated. Then, the answer from a specific objective individual is calculated into nine scores (addressing nine types of BCs). The result of BC categorization is given by identifying the highest score via winner-take-all mechanism. All the diagnosis results are referred as the reliable criterion for the prediction. This diagnosis is introduced as the pre-process of our investigation. From the pre-process results, the 851 objective individuals are categorized into nine BCs as shown in Tab. 3.1.

Table 3.1: Diagnosis results of all 851 individuals

BC	GN	QF	YA	YI	PW	WH	BS	QD	SD
#	512	41	91	23	16	23	15	96	34

Simultaneously, the life-style questionnaire with 254 items, which is originally created by our early work [59] is distributed to the same individuals. The answer

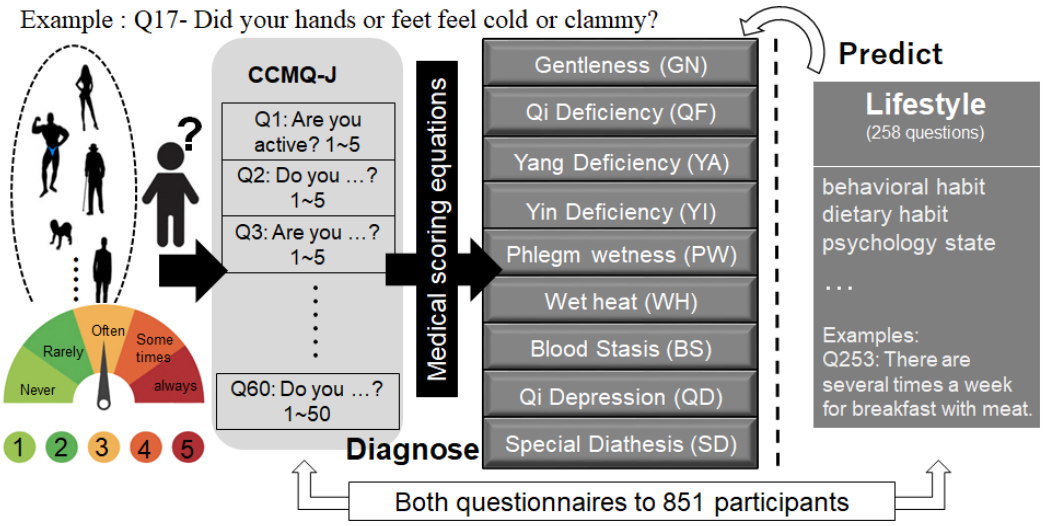

Figure 3.1: Nine types of body constitutions are diagnosed by linear scoring calculation from CCMQ-J answers; and life-style questionnaire are employed to predict BCs through machine learning algorithms.

from a specific individual is seen as a mixed vector with 254 dimensions, where both of discrete and continuous values such ages exist. Harvesting the answers from 851 individuals, the prediction is implemented by the approaches of data science. Namely, a multi-class classification problem is solved with a special consideration of principle feature (PF) extraction.

3.2 Lief-Style Questionnaire

We used crowdsourcing to recruit participants because of its diverse population, low cost and low time consuming and collected those who could read Japanese questions and write their answers, which consist of 851 answers (350 males and 501 females between 20 and 85 years old). The collection of the data was approved by the Ethics Review Committees of Nara Institute of Science and Technology and Suntory Global Innovation Center Limited. All participants agreed with the purpose of the present research via the Internet.

Initially we compared questions of CCMQ-J and those of the six questionnaires based on 7 types of attributes of traditional Chinese medical theory as

follows: (1) basic information (BI; height, weight, age, gender, etc.), (2) disease (DI; a disorder of structure or function in a human, especially one that produces dominant symptoms or that affects a specific location and is not simply a direct result of physical injury), (3) social factors (SO, habits, lifestyle, living environments, working environment, social relations, etc.), (4) mental factors (ME; cognition, tenacity and emotion, etc.), (5) dietary habits (DH; preference on foods and drinks (including the cooking style, flavor, and sided supplements), eating/drinking time, speed, relevant favors), (6) sleeping state (SL; quality, time, timing, etc.), and (7) sub-health (SH, an intermediate stage between health and disease, which is not quite either status, with no typical pathologic features). The major difference between the questions of the CCMQ-J and those of the six questionnaires is that the questions of CCMQ-J in Tab. 2.1 are mainly sub-heath (SH) questions, i.e., 55 for SH, and the others belongs to ME (2 questions), DI (2), SO (1) and 1 SL (1). In contrast, the six questionnaires consist of a variety of questions concerning DH (110), ME (52), SO (21), SH (15), SL (15), DI (4), and BI (4). Thus it can be expected that the nine body constitution scores in CCMQ-J can be explained by different factors other than SH attributes.

The full vertion of our life-style questionnaire is given in Tab. 3.2.

Table 3.2: Full version of life-style questionnaire

ID	**question contents**
Q2	SO: Life events: Death of a spouse or relatives, bankruptcy, family ceremonies, separation, divorce, occupation change, retirement, house-moving, childbirth, hospitalization.
Q5	BI: nationality
Q7	ME: K6: (Anxiety) nervous
Q8	ME: K6: (Depressed mood) hopeless
Q9	ME: K6: (Motor agitation) restless
Q10	ME: K6: (Depressed mood) depressed that nothing could cheer you up
Q11	ME: K6: (Fatigue) feel that everything was an effort
Q12	ME: K6: (Fatigue) worthless
	Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q22	ME: Job Stress: I am loaded with a lot of work.
Q23	ME: Job Stress: I cannot get the job done in time.
Q24	ME: Job Stress: I have to work very hard.
Q25	ME: Job Stress: My job requires strong powers of concentration.
Q26	ME: Job Stress: My job requires specialized knowledge and skills.
Q27	ME: Job Stress: I always have to keep my mind on the job during my duty hours.
Q28	SO: Job Stress: I am a physical worker.
Q29	ME: Job Stress: I can do my job at my own pace.
Q30	ME: Job Stress: I can decide where to start and how to do my job by myself.
Q31	ME: Job Stress: I can have my opinions reflected on the work policy at my office.
Q32	ME: Job Stress: I have a little opportunities to use my skill and knowledge in my job.
Q33	ME: Job Stress: There is a difference of opinions in my department.
Q34	ME: Job Stress: My department doesn't get along with other departments.
Q35	ME: Job Stress: My office has a comfortable atmosphere.
Q36	ME: Job Stress: The work environment of my office (noise, lighting, temperature, ventilation) is not good.
Q37	ME: Job Stress: What I am doing in my job matches me.
Q38	ME: Job Stress: My job is rewarding.
Q39	ME: Job Stress: conditions for the past month: Invigorated.
Q40	ME: Job Stress: conditions for the past month: Energetic.
Q41	ME: Job Stress: conditions for the past month: Active.
Q42	ME: Job Stress: conditions for the past month: Angry.
Q43	ME: Job Stress: conditions for the past month :Irritated.
Q44	ME: Job Stress: conditions for the past month: Annoyed.
Q45	ME: Job Stress: conditions for the past month. Exhausted.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q46	ME: Job Stress: conditions for the past month: Worn out.
Q47	ME: Job Stress: conditions for the past month: Dull.
Q48	ME: Job Stress: conditions for the past month: Tense.
Q49	ME: Job Stress: conditions for the past month: Nervous.
Q50	ME: Job Stress: conditions for the past month: Restless.
Q51	ME: Job Stress: conditions for the past month: Depressed.
Q52	ME: Job Stress: conditions for the past month: I am not up for doing anything.
Q53	ME: Job Stress: conditions for the past month: I cannot concentrate on anything.
Q54	ME: Job Stress: conditions for the past month: I am depressed.
Q55	ME: Job Stress: conditions for the past month: I am distracted.
Q56	ME: Job Stress: conditions for the past month: I feel sad.
Q57	SH: Job Stress: conditions for the past month: I feel dizzy.
Q58	SH: Job Stress: conditions for the past month: I feel a pain in my joints.
Q59	SH: Job Stress: conditions for the past month: I feel heavy in the head or have a headache.
Q60	SH: Job Stress: conditions for the past month: I have bad stiff neck and shoulders.
Q61	SH: Job Stress: conditions for the past month: I have a backache.
Q62	SH: Job Stress: conditions for the past month: I have eyestrain.
Q63	SH: Job Stress: conditions for the past month: I feel my heart pounding or I get out of breath.
Q64	SH: Job Stress: conditions for the past month: My stomach is not in good shape.
Q65	SH: Job Stress: conditions for the past month: I have no appetite.
Q66	SH: Job Stress: conditions for the past month: I suffer from diarrhea or constipation.
Q67	SL: Job Stress: conditions for the past month: I cannot sleep well.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q68	SO: Job Stress: How casually can you talk with your boss?
Q69	SO: Job Stress: How casually can you talk with your colleagues?
Q70	SO: Job Stress: How casually can you talk with your spouse, families, friends?
Q71	SO: Job Stress: When you are in a trouble, how much can you rely on your boss?
Q72	SO: Job Stress: When you are in a trouble, how much can you rely on your colleagues?
Q73	SO: Job Stress: When you are in a trouble, how much can you rely on your spouse, families, friends?
Q74	SO: Job Stress: If you talk to your boss about your personal things, how much does he listen to?
Q75	SO: Job Stress: If you talk to your colleagues about your personal things, how much do they listen to?
Q76	SO: Job Stress: If you talk to your spouse, families, friends about your personal things, how much do they listen to?
Q77	SO: Job Stress: satisfaction score: I am satisfied with my job.
Q78	SO: Job Stress: satisfaction score: I am satisfied with my home life.
Q140	ME: Personality: TIPI-J: self-assessment: I see myself as dependable, self-disciplined.
Q141	ME: Personality: TIPI-J: self-assessment: I see myself as critical, quarrelsome.
Q142	ME: Personality: TIPI-J: self-assessment: I see myself as anxious, easily upset.
Q143	ME: Personality: TIPI-J: self-assessment: I see myself as extraverted, enthusiastic.
Q144	ME: Personality: TIPI-J: self-assessment: I see myself as open to new experiences, complex.
Q145	ME: Mental: Personality: TIPI-J: self-assessment: I see myself as reserved, quiet.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q146	ME: Personality: TIPI-J: self-assessment: I see myself as sympathetic, warm.
Q147	ME: Personality: TIPI-J: self-assessment: I see myself as disorganized, careless.
Q148	ME: Personality: TIPI-J: self-assessment: I see myself as calm, emotionally stable.
Q149	ME: Personality: TIPI-J: self-assessment: I see myself as conventional, uncreative.
Q153	BI: Gender
Q155	BI: Height
Q156	BI: Weight
Q159 (1-4)	SO: Marital status: (1) No, (2) Yes, (3) separated, (4) bereaved
Q160	SO: The number of members under the same roof including you
Q163	SO: The local residents are helping each other.
Q164	SO: The local residents are reliable.
Q165	SO: The local residents exchange greetings.
Q166	SO: When a problem occurs, the local residents try to solve together.
Q167 (1-3)	SO: Educational background: (1) I graduated., (2) I am currently at school., (3) I am not at school.
Q169	SO: Did you earn money at work in the past month?
Q182	DI: Current health condition
Q183	DI: Do you have any health problems which affect your daily life?
Q185	DI: to be admitted to a hospital, dentist, acupuncturist, or massage practioner for treatments of injuries or diseases now.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q187 (1-43)	DI: (1) Diabetes, (2) Obesity, (3) Dyslipidemia(hypercholesterolemia etc.) (4) Thyroid diseases (5) Depression or other mental diseases, (6) Dementia, (7) Parkinson's disease, (8) Other nervous diseases (Neuralgia, Paralysis), (9) Eye diseases, (10) Ear diseases, (11) High blood pressure, (12) Stroke (cerebral hemorrhage, cerebral infarction etc.), (13) Angina pectoris, Myocardial infarction, (14) Other cardiovascular diseases, (15) Acute rhinitis and pharyngitis(Cold), (16) Allergic rhinitis, (17) Chronic obstructive pulmonary disease, (18) Asthma, (19) Other respiratory diseases, (20) Stomach duodenum diseases, (21) Hepatobiliary diseases, (22) Other gastrointestinal diseases, (23) Dental diseases, (24) Atopic dermatitis, (25) Other skin diseases, (26) Gout, (27) Rheumatoid arthritis, (28) Arthropathy, (29) Stiff shoulders, (30) Low back pain, (31) Osteoporosis, (32) Kidney diseases, (33) Benign prostatic hyperplasic, (34) Menopause, (35) Fracture, (36) Injuries or burns other than fractures, (37) Anemia, blood diseases, (38) Cancers, (39) Pregnancy, Puerperium (threatened abortion, Placenta previa etc.), (40) Infertility, (41) Others, (42) Unknown,(43)I have never seen a doctor, dentist, acupuncturist, or massage practioner.
Q192-210	During the past month
Q192	SL: PSQI: How long has it taken you to fall a sleep each night?
Q194	SL: PSQI: How many hours of actual sleep did you get at night?
Q195	SL: PSQI: you had trouble sleep in because you cannot get to sleep within in 30 minutes.
Q196	SL: PSQI: you had trouble sleep in because you wake up in the middle of the night or early morning.
Q197	SL: PSQI: you had trouble sleep in because you have to get up to use the bathroom.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q198	SL: PSQI: you had trouble sleep in because you cannot breathe comfortably.
Q199	SL: PSQI: you had trouble sleep in because you cough or snore loudly.
Q201	SL: PSQI: you had trouble sleep in because you feel too cold.
Q202	SL: PSQI: you had trouble sleep in because you feel too hot.
Q203	SL: PSQI: you had trouble sleep in because you have bad dream.
Q204	SL: PSQI: you had trouble sleep in because you have pain.
Q207	SL: PSQI: how would you rate your sleep quality overall?
Q208	SL: PSQI: how often you take medicine to help you sleep?
Q209	SL: PSQI: how often have you had trouble staying awake while driving, eating meals, or engaging in social activity?
Q210	ME: PSQI: how much of a problem has it been for you to keep up enthusiasm to get things done?
Q211	SH: Do you smoke?
Q213	SH: How many teeth of your own do you have?
Q214	SH: About how you chew, please choose one applicable to you.
Q215	SH: Do you eat slowly and chew well?
Q216	SH: Did you find it more difficult to eat hard things compared to six months ago?
Q217	DH: Are you sometimes choked on tea or soups?
Q218	DH: Do you feel dry in your mouth?
Q219	DH: Can you chew well with molars on the both sides?
Q222	DH: How often do you eat out?
Q223	DH: How often do you take out?
Q224	DH: Do you have an appetite?
Q226	DH: Do you have as much appetite as you had when you were young?
Q227	DH: Do you have a difficulty eating well?
Q228	DH: Do you want lose more weight?

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q229	DH: How much do you eat until you feel satisfied?
Q230	DH: Do you eat well?
Q231	DH: How many meals do you have a day?
Q232	DH: Do you want to eat more than now?
Q234	ME: Stress: Are you worried or stressed in your daily life now?
Q236	DH: How many times do you drink Chinese tea(Oolong, Jasmine, Pu're)?
Q237	DH: How much Chinese tea (Oolong, Jasmine, Pu're tea) do you drink per time?
Q238	DH: How many times do you drink water, mineral water?
Q239	DH: How much water, mineral water do you drink per time?
Q241	DH: How many bowls of rice do you eat at breakfast in a week?
Q242	DH: How many bowls of rice do you eat at lunch in a week?
Q243	DH: How many bowls of rice do you eat at dinner in a week?
Q244	DH: How many slices of bread or bowls of cereals do you eat at breakfast in a week?
Q245	DH: How many slices of bread or bowls of cereals do you eat at lunch in a week?
Q246	DH: How many slices of bread or bowls of cereals do you eat at dinner in a week?
Q247	DH: How many bowls of noodles do you eat at breakfast in a week?
Q248	DH: How many bowls of noodles do you eat at lunch a week?
Q249	DH: How many bowl of noodles do you eat at dinner in a week?
Q250	DH: How many rice dishes do you eat in a week?
Q251	DH: How many times do you eat curry rice or Hayashi rice in a week?
Q252	DH: Amount of meat or meat products you eat per time at breakfast.
Q253	DH: How many times a week do you eat meat or meat products at breakfast?

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q254	DH: Amount of meat or meat products you eat per time at lunch.
Q255	DH: How many times a week do you eat meat or meat products at lunch?
Q256	DH: Meat, Amount meat or meat products you eat per time at dinner.
Q257	DH: How many times a week do you eat meat or meat products at dinner?
Q258	DH: Amount of sea food you eat per time at breakfast.
Q259	DH: How many times a week do you eat sea food at lunch?
Q260	DH: Amount of sea food you eat per time at lunch.
Q261	DH: How many times a week do you eat sea food at lunch?
Q262	DH: Amount of sea food you eat per time at dinner.
Q263	DH: How many times a week do you eat sea food at dinner?
Q264	DH: How many eggs do you eat in a week?
Q265	DH: Amount of soy or soy products you eat per time at breakfast.
Q266	DH: How many times a week do you eat soy or soy products at breakfast?
Q267	DH: Amount of soy or soy products you eat per time at lunch.
Q268	DH: How many times a week do you eat soy or soy products at lunch?
Q269	DH: Amount of soy or soy products you eat per time at dinner.
Q270	DH: How many time a week do you eat soy or soy products at dinner?
Q271	DH: How many glasses of milk do you drink in a week?
Q272	DH: How many times a week do you take dairy products?
Q273	DH: Amount of algae you eat per time.
Q274	DH: How many times a week do you eat algae?
Q275	DH: Amount of small fish you eat per time.
Q276	DH: How many times a week do you eat small fish?
Q277	DH: Amount of green vegetables you eat per time at breakfast.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q278	DH: How many times a week do you eat green vegetables at breakfast?
Q279	DH: Amount of green vegetables you eat per time at lunch.
Q280	DH: How many times a week do you eat green vegetables at lunch?
Q281	DH: Amount of green vegetables you eat per time at dinner.
Q282	DH: How many times a week do you eat green vegetables at dinner ?
Q283	DH: Amount of light-colored vegetables and mushrooms you eat per time at breakfast.
Q284	DH: How many times a week do you eat light-colored vegetables and mushrooms at breakfast?
Q285	DH: Amount of light-colored vegetables and mushrooms you eat a time at lunch.
Q286	DH: How many times a week do you eat light-colored vegetables and mushroom sat lunch?
Q287	DH: Amount of light-colored vegetables and mushrooms you eat per time at dinner.
Q288	DH: How many times a week do you eat light-colored vegetables and mushrooms at dinner?
Q289	DH: Amount of fruits you eat a time.
Q290	DH: How many times a week do you eat fruits?
Q291	DH: Amount of potatoes you eat per time.
Q292	DH: How many times a week do you eat potatoes?
Q293	DH: Amount of jam or honey you eat per time.
Q294	DH: How many times a week do you eat jam or honey?
Q295	DH: Amount of boiled food you eat per time.
Q296	DH: How many times do you eat boiled food?
Q297	DH: Amount of food seasoned with vinegar you eat per time.
Q298	DH: How many times a week do you eat food seasoned with vinegar?
Q299	DH: How many times a week do you eat Japanese sweet?

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q300	DH: How many time a week do you eat pastry or cake?
Q301	DH: How many times a week do you eat snack food?
Q302	DH: How many pieces of rice cookies or cookies do you eat in a week?
Q303	DH: How many times a week do you eat ice cream?
Q304	DH: How many times a week do you eat candy or toffee?
Q305	DH: How many times a week do you eat jelly or pudding?
Q306	DH: How many times a week do you eat chocolate?
Q307	DH: Amount of sugar you have in a cup of coffee or tea.
Q308	DH: How many spoonful of sugar a week do you have in your coffee or tea?
Q309	DH: Amount of soft drinks or canned coffee you have at a time.
Q310	DH: How many bottles of soft drinks or cans of coffee a week do you drink?
Q311	DH: Amount of alcoholic beverages do you have per time.
Q312	DH: How many times do you drink alcoholic beverages a week?
Q313	DH: How many times a week do you have dietary supplement?
Q314	DH: Amount of butter or margarine you take per time.
Q315	DH: How many times a week do you have butter or margarine?
Q316	DH: How many times a week do you have deep-fry?
Q317	DH: How many times a week do you have mayonnaise or dressing?
Q318	DH: How many times a week do you have stir-fry?
Q319	DH: Amount of nuts(peanuts, almonds)you have per time.
Q320	DH: How many times a week do you have nuts(peanuts or almonds)?
Q321	DH: Amount of sesame seeds you have per time.
Q322	DH: How many times a week do you have sesame seeds?
Q323	DH: Amount of salty food you take per time.
Q324	DH: How many times a week do you have salty food?
Q325	DH: Amount of pickles you have per time.

Continued on next page

Table 3.2 – continued from previous page

ID	question contents
Q326	DH: How many times do you have pickles?
Q327	DH: Amount of soy or Worcester sauce you have per time.
Q328	DH: How many times a week do you have soy or Worcester sauce?
Q329	DH: How many cups of miso soups do you have in a week?
Q330	DH: How many cups of soups do you have in a week?
Q331	DH: Intake amount of soup of noodles.
Q332	DH: How many times a week do you eat the soups of noodle dishes?
Q333	DH: How you feel in taste about the dishes when you eat out

3.3 Prediction of BCs

The 254-dimensional classification is conducted by machine learning algorithms. Here, the diagnosed BC results are referenced as the teacher labels. Moreover, nine scores of each individual are also available, which contain much richer information than simple classification labels. It is obvious that not only the classification algorithms but also the regressions are feasible to carry out the prediction. On the other hand, the PF extraction is more concerned instead of simple prediction. Multiple algorithms are implemented for offering the highly reliable cross-validation. Four algorithms are applied in this work including random forest (RF), partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), and a novel pair-wise scheme of LASSO.

3.3.1 Random Forest

The random forest algorithm is one of the typical classifiers for complex data-set. However, this approach is unable to statically reduce the dimensions or identify key features. However, the RF offers very high quality of classification since all of features in the sample space are employed. Our purpose to implement RF is to verify the effectiveness of our created life-style questionnaire beyond any

methodology to extract the principle features. Namely, the RF is applied as the upper-bound of prediction correctness. Here, the RF regression is adapted by referencing nine sets of BC scores obtained in the pre-process. Nine sets of predicted scores are regressed by the life-styles. Then, the winner-take-all mechanism is applied to label a specific sample into the corresponding BC. The RF is an ensemble learning algorithm that grows a number of tree-based weak classifiers to prevent the overfitting problem. Each tree is grown by randomly selecting partial bootstrapped variables. Simultaneously, the RF reduces the predictive variance by decorrelating the individual weak classifier. After constructing the classifiers, the importance of variables can be voted by the contribution of the classification among the ensemble trees. The RF was implemented as follows:

Construction of the RF ensemble:

For $i = 1$ to B: (B is the number of individual tree)

- Draw a bootstraped sample of size n from the training data;
- Grow a tree T_i to the bootstrap sample data;
- Repeat the following steps until the minimum node size n_{min} is

  a. Select m features at random from the f variables;
  b. Decide the best feature to split the data;
  c. Spit the node using the feature selected and grow the tree to the maximum depth d.

Output the ensemble of tree $\{T_B\}$.

Ranking the feature importance:

Statistics of classification contribution from $\{T_B\}$.

The RF reduces the variance of the ensemble in accordance with the following equation:

$$Var = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3.1}$$

where the ρ is the correlation of trees and σ^2 is the variance of the features (assumed as a constant for all features). By increasing the tree numbers, the

second term on the right-hand side becomes minor; whereas the first term can be decreased by reducing the correlation of trees by a random selection of the m features.

The hyperparameters for the RF are the number of trees, max features, and max depth of a tree. Since the computation time is not a concern, the appropriate larger number of trees (200 in this study) is used for minimizing the variance. As mentioned above, only a portion of the independent features is used to grow a tree and the number is defined by the maximum variables ($m = \sqrt{f}$). The tree with a deep depth might leads to the overfitting problem, here, the max depth of a tree must be defined. ($d = 10$ in this study).

3.3.2 PLS regression

The PLS algorithm is applied to predict BCs and extract the principle features. This method has been widely used in medical imaging [34] as well as the chemo- and bio-informatics fields [18, 24, 68] since PLS models can be constructed even if there are more variables than observations. In addition, the PLS can be applied if multi-collinearities are hidden between the independent variables. Similarly to the use of RF, the PLS is employed for regressing the nine sets of scores. The BC prediction is also given by winner-take-all. The benefit of PLS is statically ranking the contribution of features. It is feasible to identify principle features somehow from the ranking after PLS training. The fundamental of PLS is briefly introduced as follow. The objective variable, Y, corresponds to the score for one of the body constitution of CCMQ-J, and the interpretive variables $X_1, X_2, \ldots, X_M$ corresponds to answer for questions of life-style questionnaire are correlated by a linear model as:

$$Y = a_0 + a_1 X_1 + ... + a_j X_j + ... + a_M X_M. \tag{3.2}$$

Here M represents the total number of the questions.

The PLS model is represented in the following equations:

$$\mathbf{y} = \bar{\mathbf{y}} + \sum_{k=1}^{A} \mathbf{t}_k q_k + \mathbf{e} = \bar{\mathbf{y}} + \mathbf{T} \cdot \mathbf{q} + \mathbf{e}. \tag{3.3}$$

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^{A} \mathbf{t}_k {\mathbf{p}_k}^T + \mathbf{E} = \bar{\mathbf{X}} + \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E}, \tag{3.4}$$

where q_k is the coefficient of y for the k^{th} component, p_k is the loading vector of X, A is the number of components (features), and t_k is a score vector for the k^{th} component. The residual matrix and vector are represented by $E(M \times N)$ and $e(M \times 1)$, respectively. Eqs (3) and (4) can be combined to create Eq (5).

$$\mathbf{Y} = \bar{\mathbf{y}} - \bar{\mathbf{X}}^T \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} + \mathbf{X}^T \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q}. \tag{3.5}$$

The number of PLS components was determined by maximizing the Q^2, which was calculated by a leaving-out-one cross-validation for each component, as shown in Eq (6).

$$Q^2 = 1 - \frac{\sum_{i=1}^{N} \left(y^{(i)} - y_{cv}^{(i)} \right)^2}{\sum_{i=1}^{N} \left(y^{(i)} - \bar{y} \right)^2}. \tag{3.6}$$

Here, y and $y_c^{(i)}v$ are original and predicted y-values in the cross-validation for every ith individual, respectively and $\bar{y}$ represents the average for all y-values. We determined the number of components so that Q^2 value reaches the maximum. Then after determining the number of components, we also calculated the R^2 for examining prediction accuracy for the PLS model. Those selected components are ranked by the contribution to the prediction in terms of the coefficient. It is obvious that the PLS cannot directly identify the principle features but offer the importance of each feature. A select rule is necessary, top 5% for instance.

3.3.3 LASSO Regression

In machine learning and data science, the parsimony of statistical models is crucial for exploring the proper features. The RF is a non-parametric model that feature selection is limited by the construction of ensemble trees, and the accuracy of classification determines the confidence level since the importance is ranked by the classification results. Additionally, the combinatorial search of the classical feature selection algorithm brings the heavy computation cost and the limitation of balance between the optimal model and its features.

The LASSO is a statistical method within the linear regression situation which is used for feature selection via shrinking estimates of irrelevant features towards zero. The features resulting from the regularization of LASSO construct a sparse model that provides more interpretable factors of data itself. Moreover, the entire regularization path can be computed in the complexity of one linear regression, provides a computationally feasible way for the model selection [10, 17, 51, 56, 67]. Assuming that the constitution data of lifestyle is fitted to a linear regression model with following the equation:

$$Y_n = X_n\beta^n + \varepsilon_n \tag{3.7}$$

where $\varepsilon_n = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is a bias vector of i.i.d random variable with $\varepsilon_n \sim N(0, \delta^2)$. Y_n represents the regression response corresponding to the data X_n that $X_n = (x_1^n, \cdots, x_p^n)$ is the $n \times p$ design matrix. The β^n is a vector as the model coefficient, that is, some of the regression coefficients β^n are exactly zero corresponding to the predictors that are irrelevant to the response. Unlike classical fixed p settings, the data and model parameters β are indexed by n to allow them to change as n grows, and the auxiliary condition for β allows for penalizing the absolute size of the regression coefficient on basis of the tuning parameter value λ (2) in LASSO [51]. Hence, the Lasso estimates $\beta^n = (\beta_1^n, \cdots, \beta_j^n, \cdots)^T$ are defined by

$$\hat{\beta}^n(\lambda) = argmin \left\| Y_n - X_n\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1, \tag{3.8}$$

where $\left\| \cdot \right\|_1$ stands for the L_1 norm of a vector which equals the sum of the absolute values of the vector's entries. The non-negative parameter λ controls the amount of regularization applied to the estimate, since the bigger value of λ will enhance the penalty and shrink more non-zero β. In general, moderate values of λ will cause shrinkage of the solutions towards 0, and some coefficients may end up being exactly 0 (the number of λ is 100 in this study).

Note-worthily, the optimal penalty of λ is evaluated by k-fold cross validation and the mean cross-validated error (a vector of the length of λ) determines a generalized model. Moreover, the mean cross-validated error provides a flexible range that the largest value of lambda within one standard error of the minimum represents the simplest model. To pick over the importance, the feature shrinkage

in this study is resulted from the simplest model and implemented on the 10-fold cross validation. Here, the score regression and winner-take-all are applied for prediction. Differing from the PLS, the LASSO is feasible to completely exclude non-principle features rather than a ranking result.

3.3.4 Pair-Wise Classification via LASSO

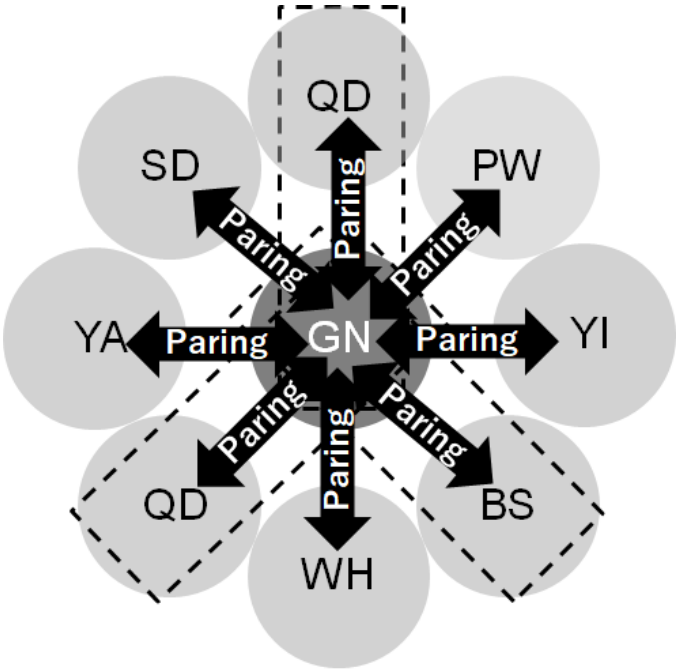

Figure 3.2: BCs are grouped into eight pairs according to CCMQ-J diagnosis results. Prediction is directly conducted by two-class classification (also known as binary logic regression).

By using all three algorithms above, the regression of BC scores are necessary for the prediction since the nine-class classification problem is difficult to solve in a high accuracy. However, our target application is not a straightforward multi-class classification task due to the following properties: (1) the sample data is pre-processed by CCMQ-J diagnosis, which indicates the BC categorization is well labeled in advance. (2) Only distinguishing the GN against other eight biased BCs is concerned for health guidance; but the transfer among biased BCs

is meaningless. Therefore, a pair-wise strategy is applied as shown in Fig. 3.2. Nine BCs are grouped into eight pairs with the common class of GN. Through the two-class classification (also known as the binary logic regression) algorithms, the prediction is expected to realize with high accuracy. As our serious concern, the principle features are identified along with the classification. Namely, it is feasible to investigate key issues in life-style leading the individual to a specific biased BC from GN. Then, the healthcare guidance can be offered for recovering to GN by changing those identified life-styles. In this work, the LASSO is employed to classify all pairs after paring.

4 Experiment Results of Prediction

The RF, PLS, LASSO algorithms are applied for regression of BC scores. The proposed pair-wise LASSO is applied by directly classifying the biased BCs against GN constitutions.

4.1 RF regression

Figure 4.1 illustrates the BC score regression for nine types by using the RF algorithm and full of 254 life-style items. From the winner-take-all result, the correct rate of BC prediction is 88.7%. Since all of features are employed in the regression, the RF algorithm gives a high accuracy on the prediction. Although the principle features can not be identified, the well performance over regression and final prediction results indicates: (1) the BCs can be precisely predicted by life-styles; (2) our created life-style questionnaire is sufficient for prediction; (3) an upper-bound of accuracy of 88.7& (or even higher) is guaranteed by using advance machine learning algorithms.

The RF algorithm appears powerful regression ability as above. It is noticed that RF also gives the important ranking somehow. However, the PF model is not interpretable since RF is a statistics model. Figure 4.2 illustrates the poor interpretability of RF model. Two trials of RF training offers completely different PFs ranking even if the final prediction accuracies are same. Thus, it is infeasible to observe the stable PFs from the learning results.

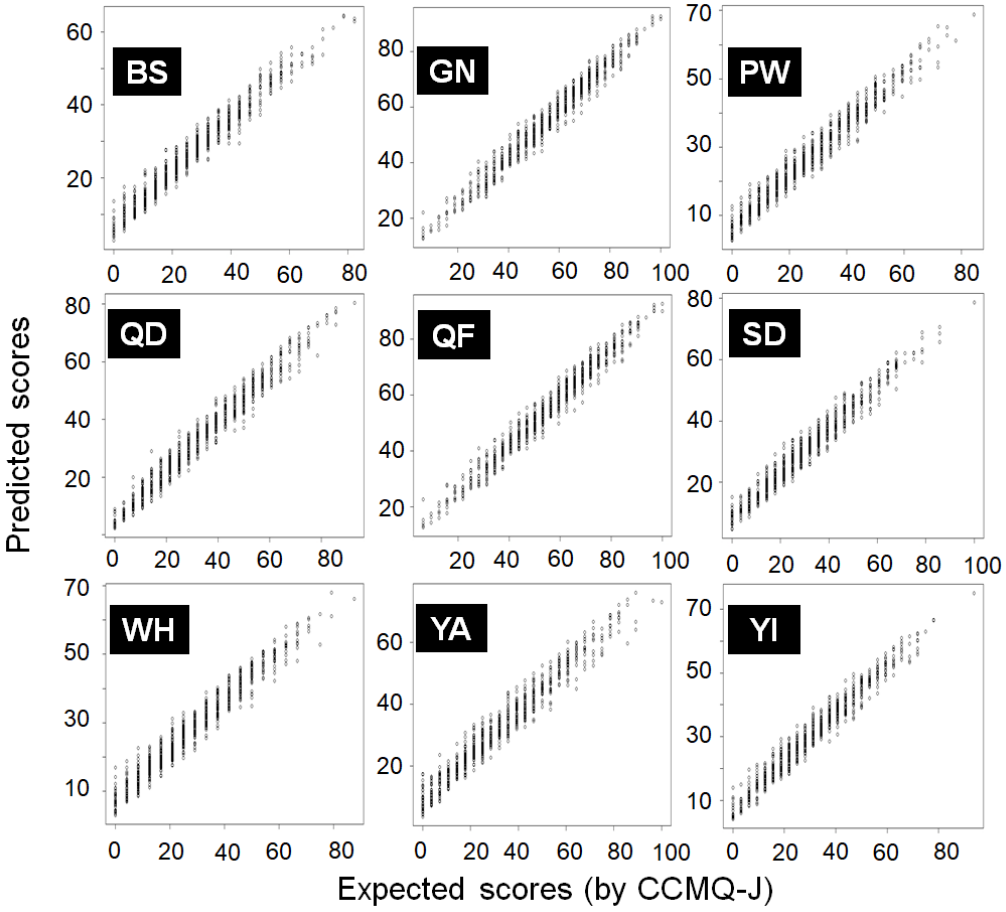

Figure 4.1: BC scores predicted by life-style questionnaire through RF algorithm against actual scores calculated by CCMQ-J

4.2 PLS Regression

The PLS method has been widely used in medical imaging as well as the chemo- and bio-informatics fields because PLS models can be constructed even if there are more variables than observations. In addition, this method can be applied if multi-collinearities are hidden between the independent variables.

	Variables	Importance	Rank
Q62	Q62	0.009965707	# 1
Q218	Q218	0.009956234	# 2
Q201	Q201	0.009529944	# 3
Q66	Q66	0.009283644	# 4
Q60	Q60	0.008857354	# 5
Q45	Q45	0.008734204	# 6
Q53	Q53	0.008364752	# 7
Q47	Q47	0.008260548	# 8
Q64	Q64	0.008241602	# 9
Q11	Q11	0.008118452	# 10
Q9	Q9	0.007569012	# 11
Q46	Q46	0.007540592	# 12
Q202	Q202	0.007445861	# 13
Q59	Q59	0.007417442	# 14
Q63	Q63	0.007389023	# 15
Q153	Q153	0.007360603	# 16
Q207	Q207	0.007303765	# 17
Q52	Q52	0.007237453	# 18

Trial 1

	Variables	Importance	Rank
Q66	Q66	0.010767050	# 1
Q64	Q64	0.009004343	# 2
Q47	Q47	0.008849879	# 3
Q201	Q201	0.008840793	# 4
Q218	Q218	0.008686329	# 5
Q59	Q59	0.008522779	# 6
Q62	Q62	0.008522779	# 6
Q202	Q202	0.008504607	# 7
Q67	Q67	0.008150248	# 8
Q11	Q11	0.007804976	# 9
Q45	Q45	0.007668684	# 10
Q65	Q65	0.007550564	# 11
Q53	Q53	0.007541478	# 12
Q207	Q207	0.007541478	# 12
Q7	Q7	0.007450617	# 13
Q52	Q52	0.007432445	# 14
Q60	Q60	0.007368842	# 15
Q10	Q10	0.007232550	# 16

Trial 2

Figure 4.2: Two trials of RF training offers completely different PFs ranking even if the final prediction accuracies are same.

4.2.1 Performance of regression

Multivariate linear regression models of 9 body constitution scores in CCMQ-J based on scores of the 254 questions in the six questionnaires (life-style questionnaire) were built by PLS. We selected the optimal number of components by Q2 values for the 9 regression models, respectively, i.e., 3 components for all body constitutions except QD and SD scores (2 and 4 components, respectively; Fig. 4.3). The relationships between the true scores according to the CCMQ-J and the predicted scores with the PLS models are represented in Fig. 4.4.

Then we constructed the 9 regression models. Figure 4.5 shows R2 values and Pearson coefficients between the original and predicted scores for the nine physical constitutions. High correlations between original and predicted scores were obtained in GN ($r = 0.90$), QD (0.86) and QF (0.86), and those of others are all higher than 0.70. Thus, it should be noted that the nine physical constitutions scores can be estimated by the scores of the questions in the six questionnaires.

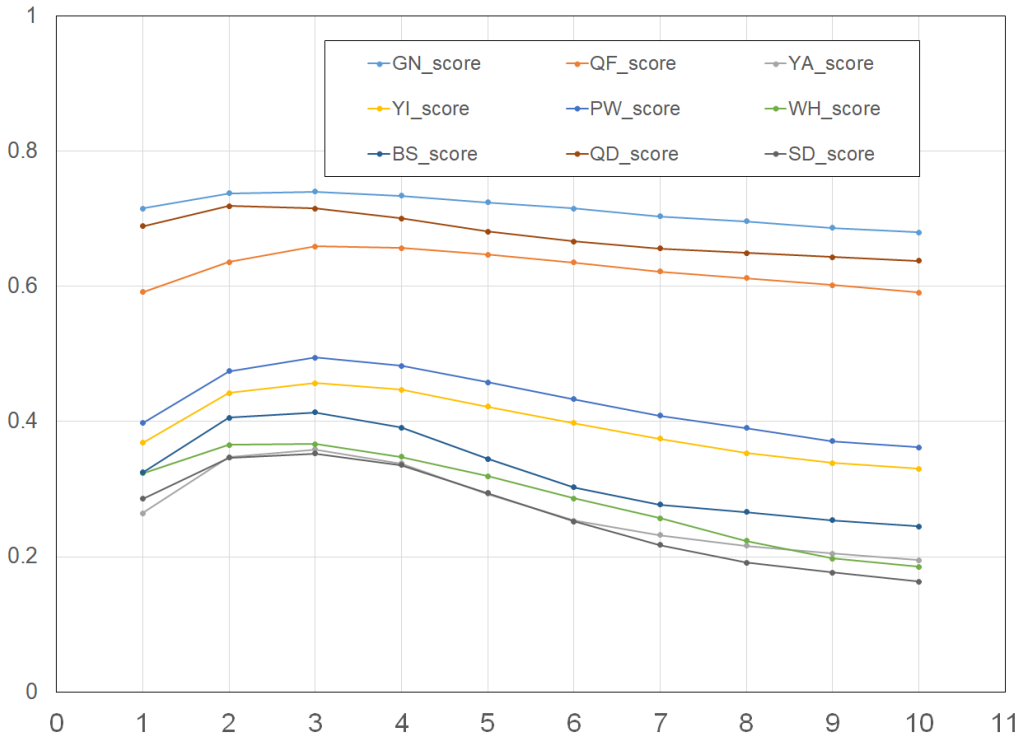

Figure 4.3: Selection of the component number for 9 scores of CCMQ-J according to the trend of Q2

Table 4.1 lists the questions with the highest 5% coefficients (both positively and negatively) in all 9 body constitutions. In average for the 9 regression models, there are 14 (from 3 to 33), and 13 (from 4 to 20) questions contribute positively and negatively, respectively.

Figure 4.6 shows the histogram of the questions in Table 3 according to the number of related body constitutions. Figure 4.7 shows the effects of the questions with the highest 5% coefficients in terms of their categories. Though most of the questions are associated with single body constitutions, 16 questions are also associated with body constitutions larger than or equal to 5.

Figure 4.8 shows the effects of the questions with the highest (positively and negatively) 5% coefficients in terms of the seven types of attributes of traditional Chinese medical theory. Y-axis represents the number of questions with the highest 5% coefficients. Negative coefficient corresponding to Fig. 4.8 (a) means

Table 4.1: Principle features from life-style questionnaire identified by PLS, LASSO, and pair-wise classification through LASSO. Question IDs are listed (full question list is seen in Tab. 3.2).

		Negative		Positive	
	Total	# of QID	QID	# of QID	QID
GN	37	33	Q7, Q9, Q10, Q11, Q12, Q45, Q46, Q47, Q50, Q51, Q52, Q53, Q54, Q55, Q60, Q62, Q67, Q142, Q145, Q147, Q149, Q164, Q182, Q195, Q196, Q197, Q201, Q207, Q209, Q210, Q228, Q293, Q306	4	Q41, Q148, Q156, Q143
BS	27	12	Q148, Q155, Q156, Q159_1, Q187_1, Q217, Q218, Q230, Q278, Q289, Q304, Q320	15	Q57, Q58, Q59, Q60, Q61, Q62, Q63, Q141, Q142, Q147, Q153, Q199, Q201, Q202, Q204
SD	28	13	Q35, Q37, Q38, Q159_3, Q185, Q187_1, Q187_34, Q217, Q218, Q242, Q311, Q312, Q319	15	Q57, Q59, Q60, Q62, Q63, Q64, Q66, Q142, Q187_15, Q187_16, Q187_18, Q187_24, Q198, Q199, Q201
QD	11	3	Q41, Q148, Q218	8	Q10, Q44, Q49, Q50, Q51, Q54, Q56, Q142
WH	35	19	Q32, Q33, Q37, Q70, Q76, Q77, Q78, Q187_12, Q187_19, Q214, Q217, Q218, Q228, Q247, Q252, Q267, Q282, Q290, Q319	16	Q43, Q44, Q60, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q201, Q202, Q219, Q224, Q237, Q306
PW	30	10	Q32, Q68, Q140, Q214, Q216, Q217, Q218, Q228, Q231, Q280	20	Q11, Q28, Q57, Q58, Q60, Q61, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q156, Q199, Q201, Q202, Q210, Q211, Q237
YI	24	11	Q32, Q78, Q155, Q156, Q160, Q187_33, Q217, Q218, Q252, Q253, Q328	13	Q57,Q59, Q61, Q62, Q63, Q66, Q142, Q199, Q64, Q60, Q153, Q201, Q202
YA	36	18	Q8, Q43, Q155, Q156, Q187_1, Q187_11, Q216, Q218, Q224, Q227, Q232, Q253, Q257, Q258, Q259, Q309, Q316, Q322	18	Q57, Q59, Q60, Q61, Q62, Q63, Q64, Q66, Q142, Q153, Q187_16, Q187_9, Q198, Q199, Q201, Q202, Q228, Q306
QF	26	11	Q31, Q40, Q143, Q187_25, Q187_42, Q217, Q218, Q224, Q227, Q317, Q333	15	Q47, Q57, Q59, Q60, Q61, Q62, Q63, Q64, Q65, Q66, Q145, Q182, Q183, Q198, Q228

that the question prevents the body from turning into the corresponding body constitution. Many questions belonging to mental factor (ME) are inhibitory for gentleness type (GN). In contrast, many sub-health (SH) questions contribute

positively to 6 types of body constitutions Blood stasis (BS), Yin-deficiency (YI), Wet-heat (WH), Phlegm-wetness (PW), Yang-deficiency (YA) and Qi-deficiency (QF) and disease factors (DI) are also associated with YA and QF (Fig. 4.8 (b)).

The principle features counted as illustrated in Fig. 4.4 are selected by the top 5% significance for each BC since the PLS algorithm hardly excludes the insignificant features completely. In this manner, the average number of PFs is reduced to 28. By winner-take-all, the correct rate of prediction is 40.9%.

4.2.2 Medical analysis

Top ten key-questions

We obtained regression models for the nine body constitutions from the scores for the 254 questions of the six questionnaires 10-20). Those regression models can provide interpretation for the nine body constitutions from multifaceted viewpoints of lifestyle in sense of medical analysis. Table 3 lists the questions with the top 5% highest weights on any constitution (both positively and negatively). The top ten key-questions are ranked as follows along with their most relevant body constitutions; Q63 for QF (0.132), Q187_16 for SD (0.106), Q201 for YA (0.100), Q57 for QF (0.095). Q66 for YI (0.090), Q218 for YI (-0.087), Q228 for PW (-0.084), Q62 for YI (0.080), Q156 for YA (-0.076) and Q153 for BS (0.075).

Q63 (SH, "I feel my heart pounding or I get out of breath") associates with the fact that palpitation and asthma are typical symbols of hypo-functioning of hearts and lungs, which indicates the weakness in blood circulation. From TCM theory, they are very strong evidence for QF, which leads to "Speaking weakly, sweating easily, shortness of breath, often tired and weak." 26). This statistical result correctly matches the clinical diagnosis.

Q187_16 (DI, Allergic rhinitis) reflect SD constitution, in which symptoms related to asthma, itching of pharynx, nasal congestion, and sneezing are frequently seen.

Q201 (SL, trouble of sleep in because of feeling of too cold) is medically categorized into YA constitution with typical clinical manifestations of chilly feeling on hands and feet, antagonism to cold foods. The chilly feel is physically caused by the decrease in metabolic rate; greater caloric dissipation than the production.

Q57 (SH, feel dizzy) could be caused by many complicated reasons, in which the central nervous system injury caused by peripheral vestibular dysfunction and trauma is not considered in this study. Naturally, the dizziness is often seen as the result of the insufficiency of blood supply in the brain. This manifestation is due to (1) excessive oxygen consumption in the brain, mental tension or sustained excitement, which may lead to WH constitution; and (2) weakness of blood circulation, obstruction of arterial blood supply due to cardio-cerebral vascular diseases or cervical vertebra deformation, which may lead to QF constitution.

In Q66 (SH, I suffer from diarrhea or constipation), diarrhea or constipation are both gastrointestinal dysfunctions due to water imbalance. In Chinese medical theory, the inadequacy of water is a typical symbol of YI constitution. pieces of evidence are seen as fear of heat, fever in the hands and feet, flushing or redness on the cheeks, dry skin, dry mouth and being prone to be insomnia, and dry stools.

Q218 (feel dry in mouth) is also associated with water inadequacy as well as Q66.

As natural comprehensions, the wish of "lose more weight" corresponding to Q228 (DH) indicates the over-weight problem, which is the most obvious evidence of PW constitution. The manifestations are: the abdomen is soft and fatty; the skin is oily, sweaty; the eyes are swollen, and sleepy.

The TCM theory points out that "over-use of eyes is harmful to blood circulation" which corresponds to Q62 (SH, eyestrain) and that "liver performs the storage and balance of blood". Thus, the eyestrain indicates the deficiency of liver and blood leads to the dehydration in skin, mouth, stools, and even insomnia, which are typical evidence of YI constitution.

Q156 (BI, Weight (Body-Mass)) is relevant to body constitutions in many senses. For instances, obesity is positively correlated to QF and PW, while negatively correlated to YA, BS, and QD. However, the relation between weight and YA is still unknown in medical science 27). From the statistical results, the weight index strongly impacts YA prediction.

It has been clearly reported that the males and females appear different body constitutions in physiologic pathology. For instance, males incline towards higher risks of severe respiratory infection, central nervous virus infection, viral gas-

troenteritis and hepatitis. However, there is no well proven theory explaining that gender (Q153, BI) particularly impacts BS body constitutions.

In summary, most of the very-high-valued Key-questions can be perfectly or indirectly explained by the medical theory. However, several ones such as obesity and weight are difficult to match to the existing clinical theory in details. There is an open potential to explore the sophisticated mechanisms on the basis of statistical results from this paper. Ranking in order of appearance-counting

Ranking in order of appearance-counting

Some questions are identified as "key" in the prediction of not only one but multiple body constitutions (Fig.s 4.6 and 4.7). Obviously, the frequency of appearance reflects the generality of a specific question. This fashion of ranking offers a different point of view for medical analysis. Not only the absolute impact of one specific condition/lifestyle but also the generality should be concerned. For instance, Q218 ("Do you feel dry in your mouth?") which reflects the water inadequacy is effected for the eight body constitutions except for GN type. It is well accepted highly advocated that the water balance should be paid more attention in the clinical treatment. In contrast, this analysis shows some features that are different from current medical theory, which might inspire some re-considerations on the medical side. For instance, the age factor is widely and strongly considered as a general factor on the body constitution prediction. This is due to the senescence of organs is irreversible and unavoidably during human life. However, from the statistical results, age is not a universal key factor. In this sense, there is a possibility (or even evidence) to compensate for the aging through lifestyles such as those factor in Fig. 4.7.

Categorization in the medical sense

Several medical implications are found based on the body constitution investigation. Firstly, the diseases (DI) are widely considered as one of the strongest signs of health status as they are reflected in YA and QF (Fig. 4.8). On the other hand, the mental factors (ME) have the most significant impacts on the body constitution balance, especially, several questions of ME are inhibitory for GN.

On the basis of above analysis, some medical explanations are made as: (1) for the body constitution balance, mental management (ME) is the most significant, even feasible to compensate for the effects of aging (BI) and disease (DI); (2) mental management (ME) strongly impacts each of the body constitutions as either positive or negative factors, which means that each body constitution manifests explicit mental feature. (3) It has been indicated by the medical theory that the impacts of sleep state (SL) and dietary state (DH) are complementary, which means the collaborative mediation is an efficient manner to escape from/drop into a specific body constitution.

Gentleness Constitution

Among the nine body constitutions, GN is especially noticeable in terms of both statistics and medical science. From the medical point of view, GN is considered as an expected "healthy" status; and all other eight body constitutions are considered as "biased" constitutions. From our statistical results (Table 3 and Figures 6), it is also found that GN shows very different features from the others and leads to some noticeable medical hints, i.e., (1) The number of negative factors for GN is much more than the positive ones (Figure 6). (2) Mental and sleep relevant issues have considerable influence in making individuals escape from GN to biased body constitutions; in contrast, it is difficult to turn back to GN by only mental and sleep efforts; and (3) in general, mental status is most significant to distinguish GN against the others.

Summary

The statistical results from our proposed method are correctly explainable for the nine types of body constitutions in multifaceted viewpoints of health conditions and understandable through existing medical theory; moreover, new medical hints, which have not been explicitly indicated by traditional medical researches, have been extracted from our data analysis. Crowdsourcing makes it possible to accumulate data from the large network of potential participants and bring people together to harness their collective information. Therefore, it is an efficient methodology to verify an important scientific assumption.

4.3 LASSO regression

The LASSO regression of scores for nine BCs are shown in Fig. 4.9 along with the number of identified PFs for each BC. The correct rate 69.9% of prediction is achieved by the winner-take-all. Both of regression and prediction performances are better than the PLS algorithms and worse than the RF. The number of PFs is not constant but flexible by changing *Lambda* parameter (see above section). Here, the demonstrated number of PFs are selected by the optimized trade-off between accuracy (evaluated by the mean squared errors of regression) and number of PFs. Figure 4.10 illustrates this trade-off map.

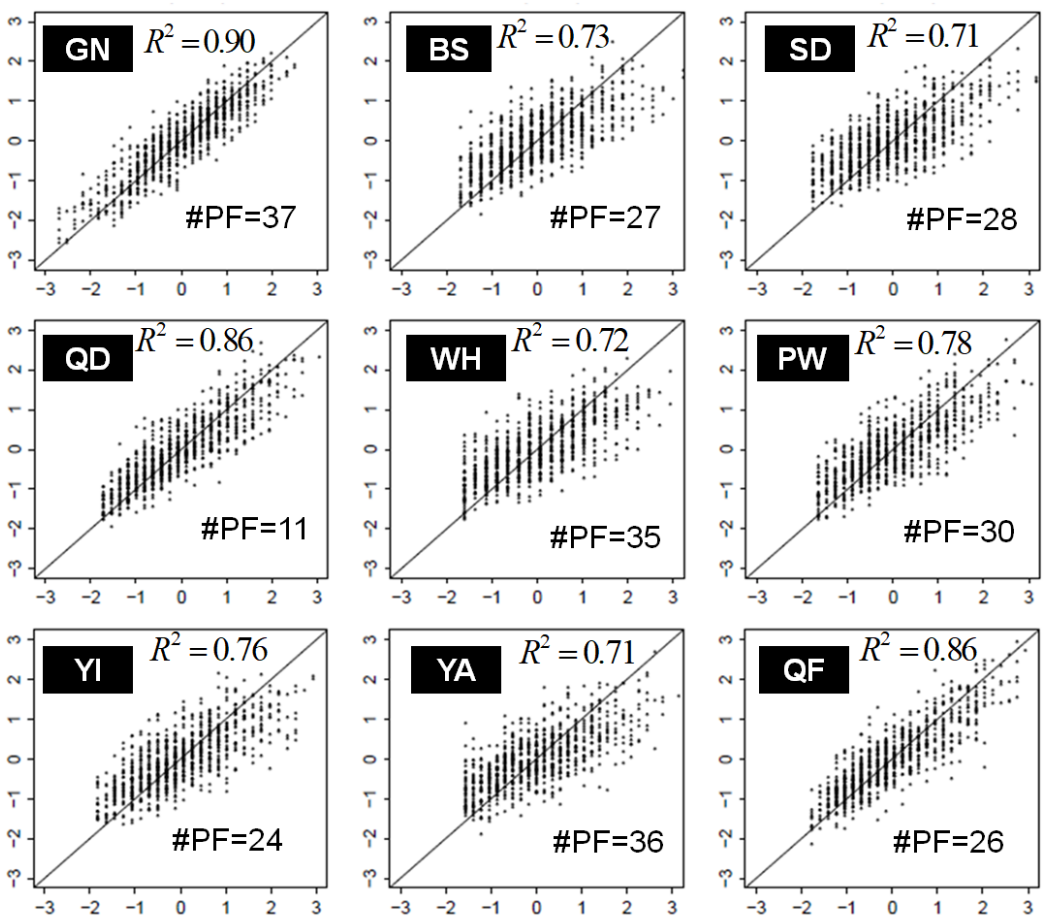

Figure 4.4: BC score regression is evaluated by Pearson coefficients for each BC with maximum R^2 values. Principle features are selected by top 5% significance.

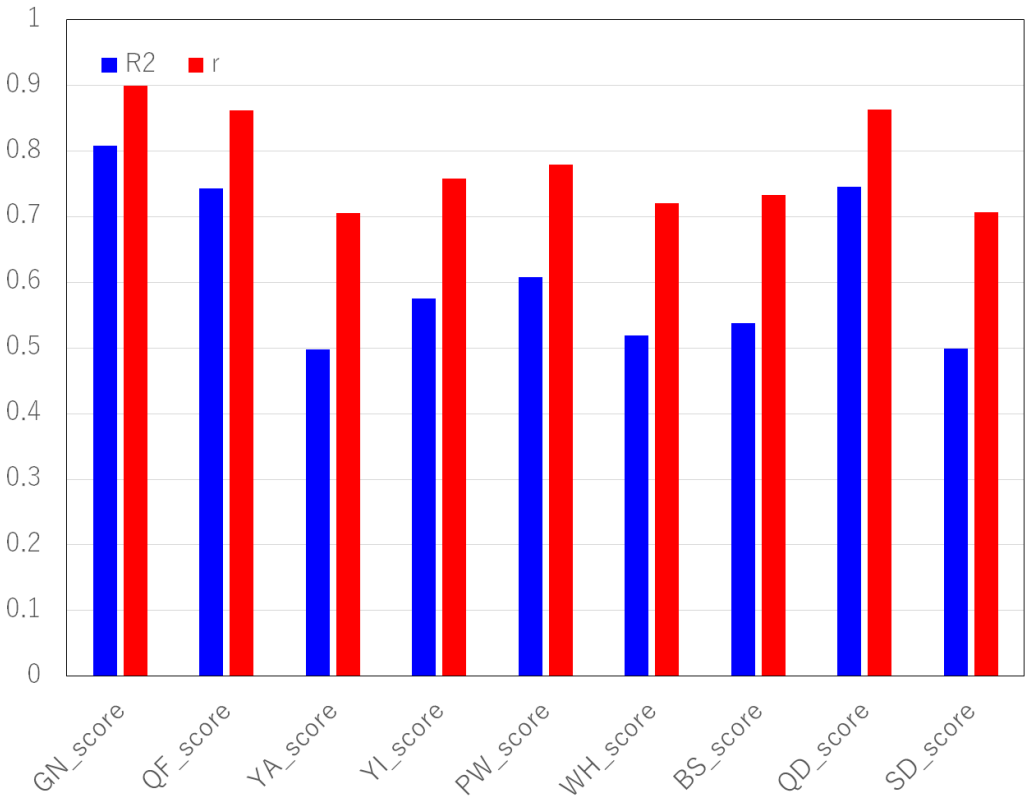

Figure 4.5: Prediction ability for PLS models in terms of the R2 and correlation coefficients (r)

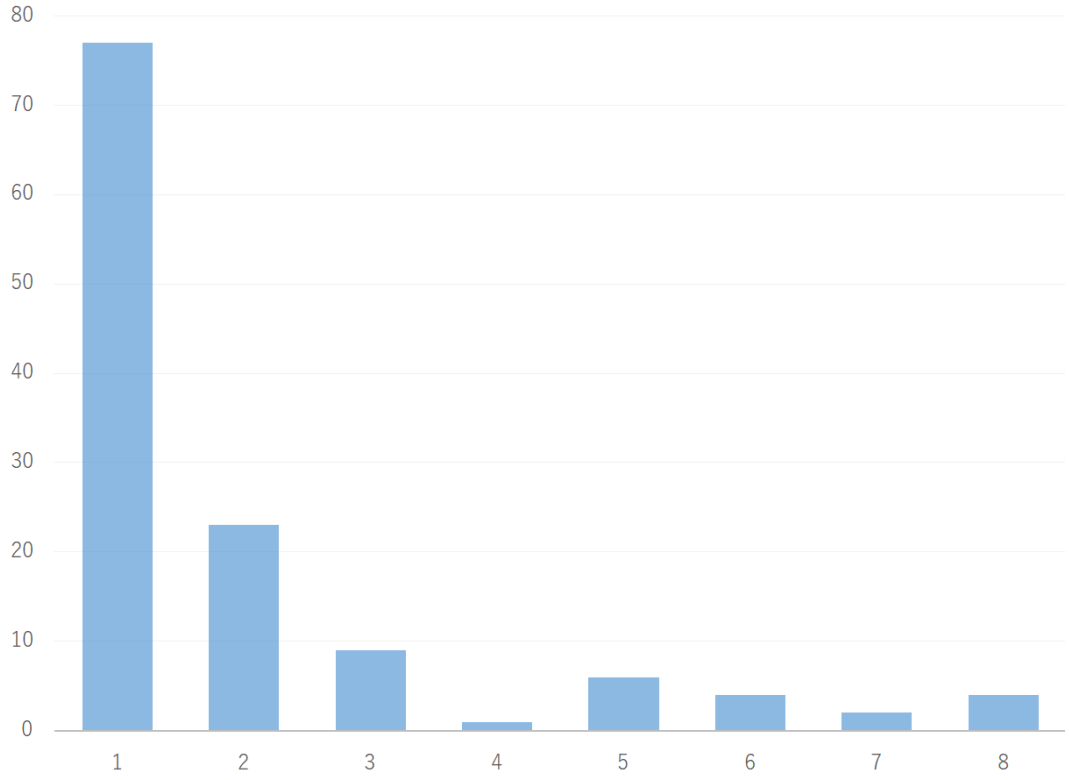

Figure 4.6: Relationship between the number of body constitutions (X-axis) and the number of questions (Y-axis) with the top 5%significances

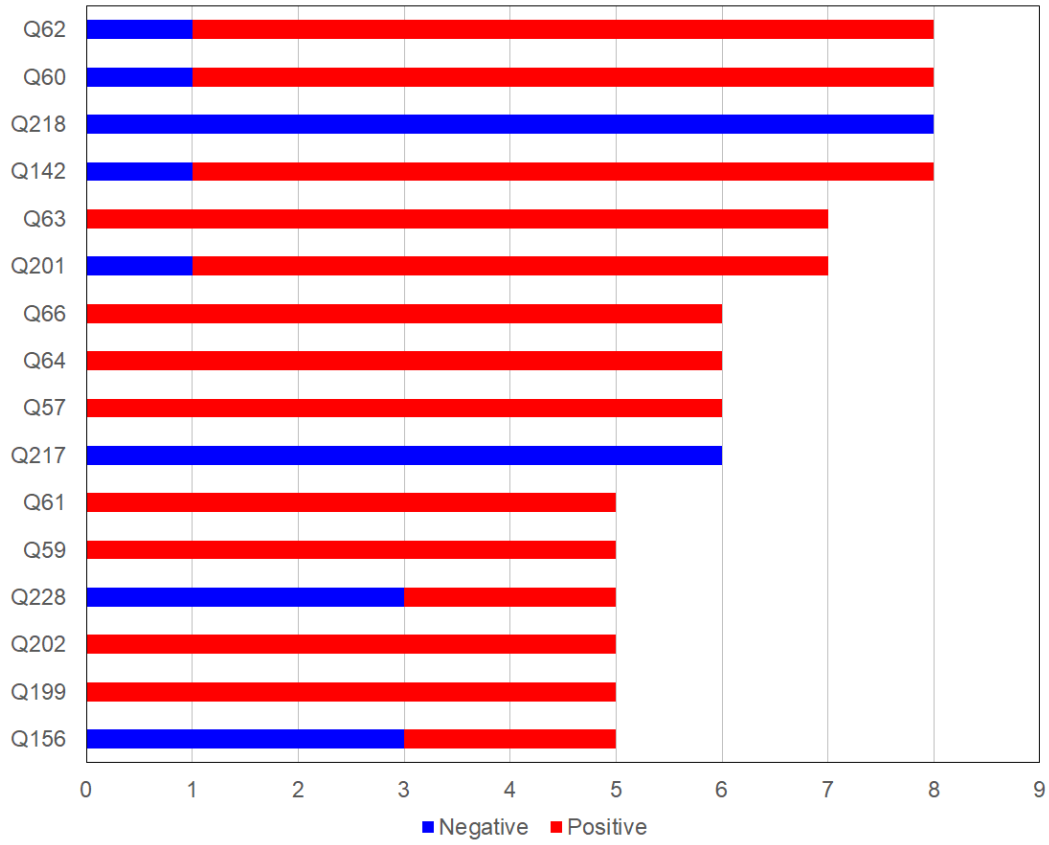

Figure 4.7: Frequency of appearance of questions with the 5% significances

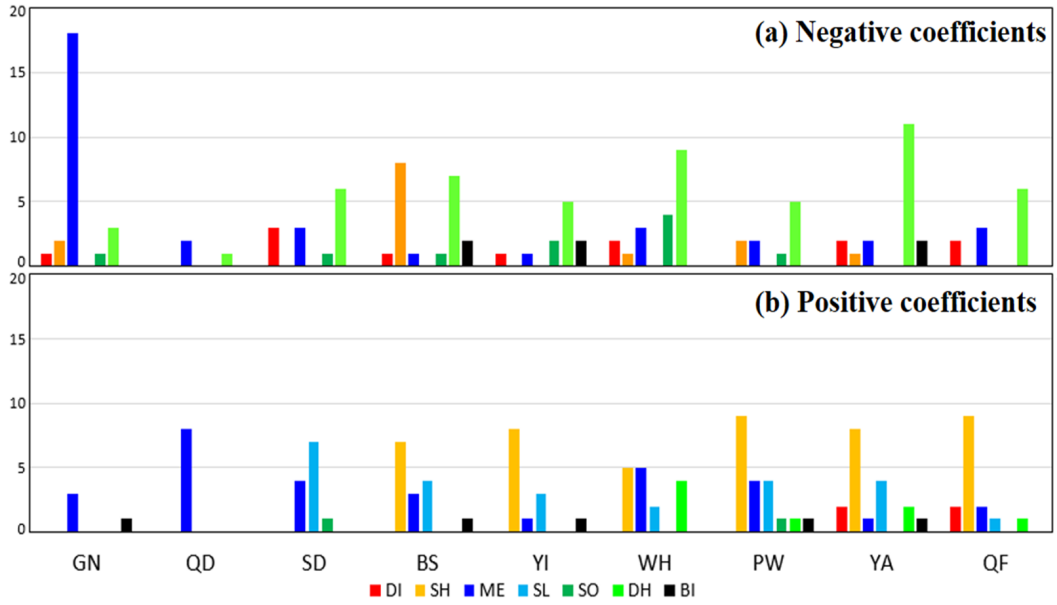

Figure 4.8: Frequency of appearance of questions with the 5% significances that have been assigned to the 7 categories (7 categories are explained in the text.)

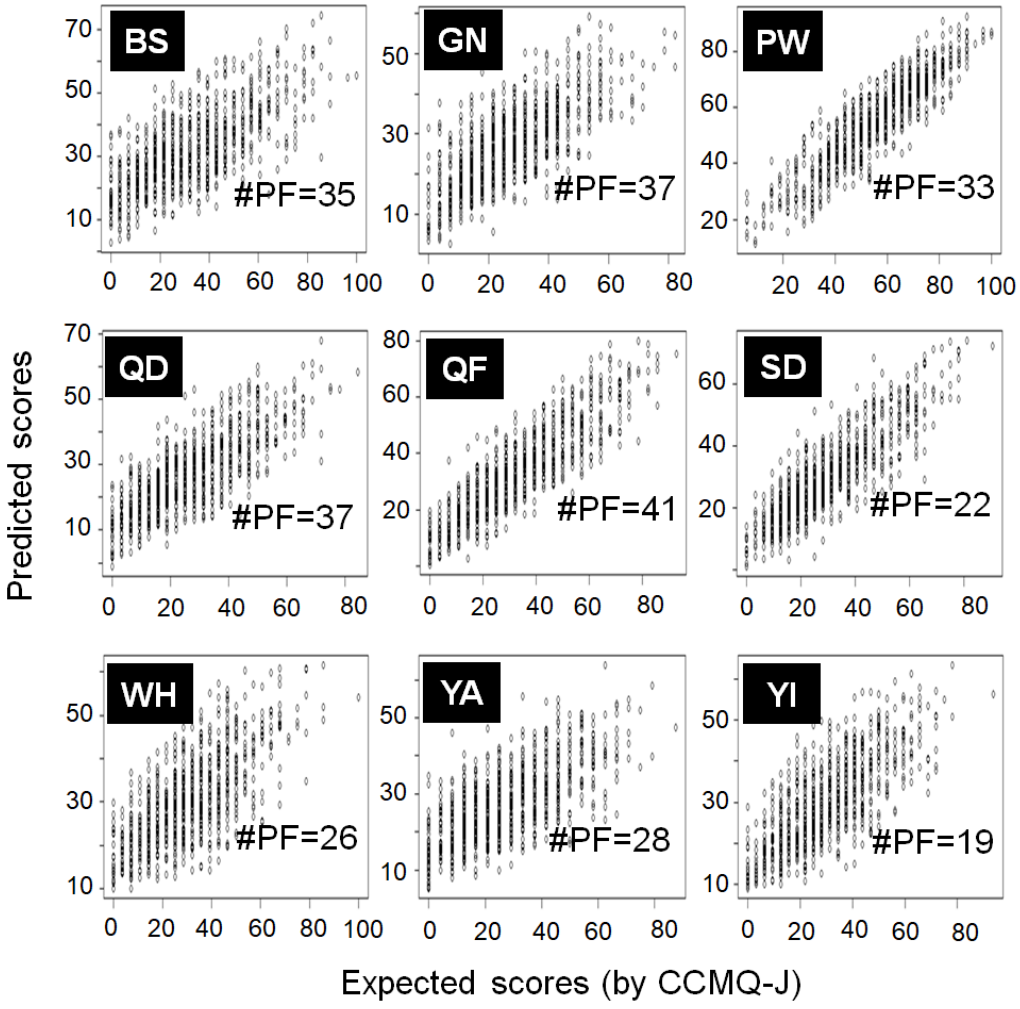

Figure 4.9: BC score is regressed by LASSO. Principle features are selected by identifying features with non-zero coefficient in trained LASSO model.

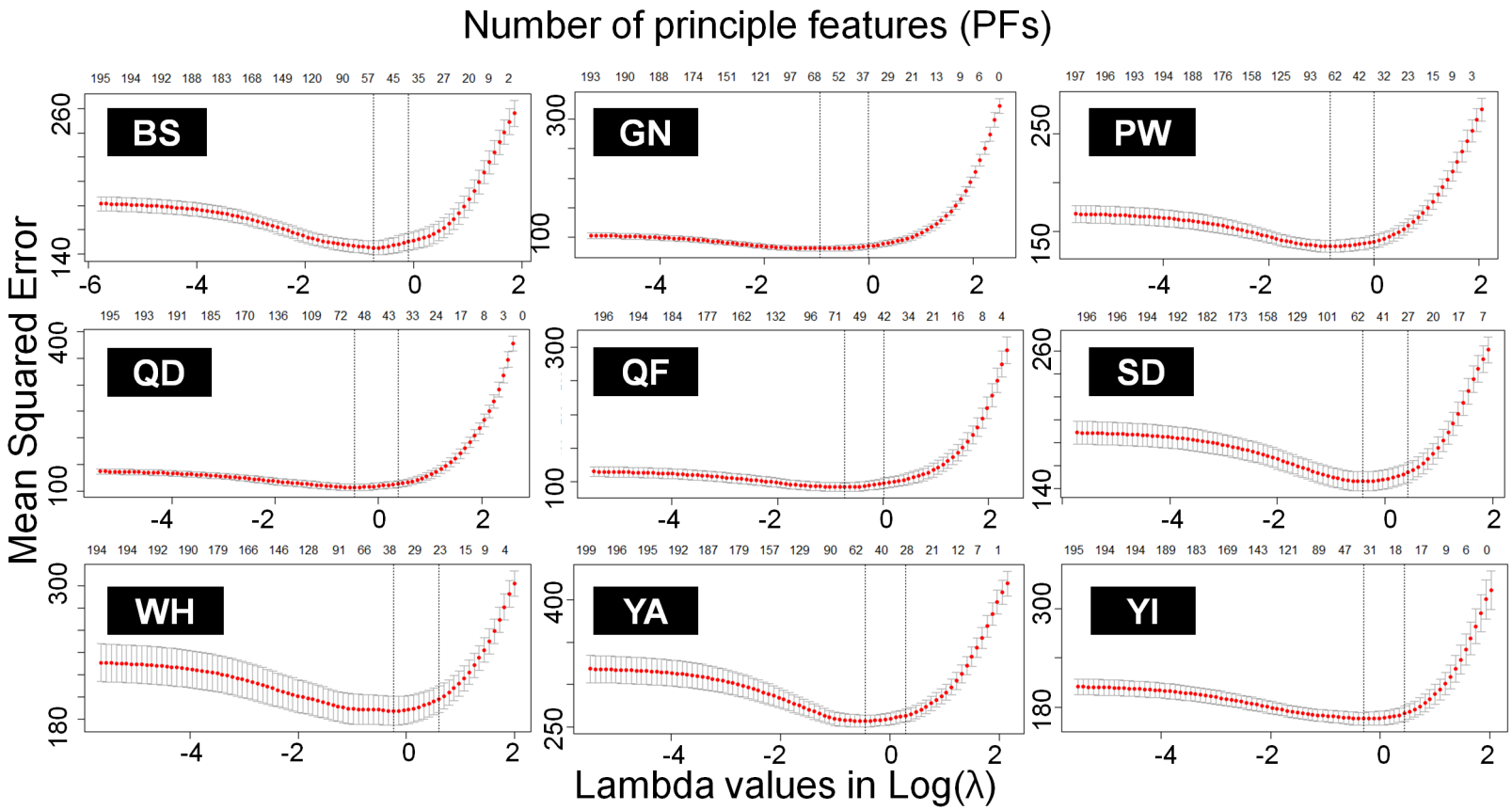

Figure 4.10: Number of PFs is traded for high regression quality by adapting different *Lambda* values. Optimum range of *Lambda* is obtained by LASSO training.

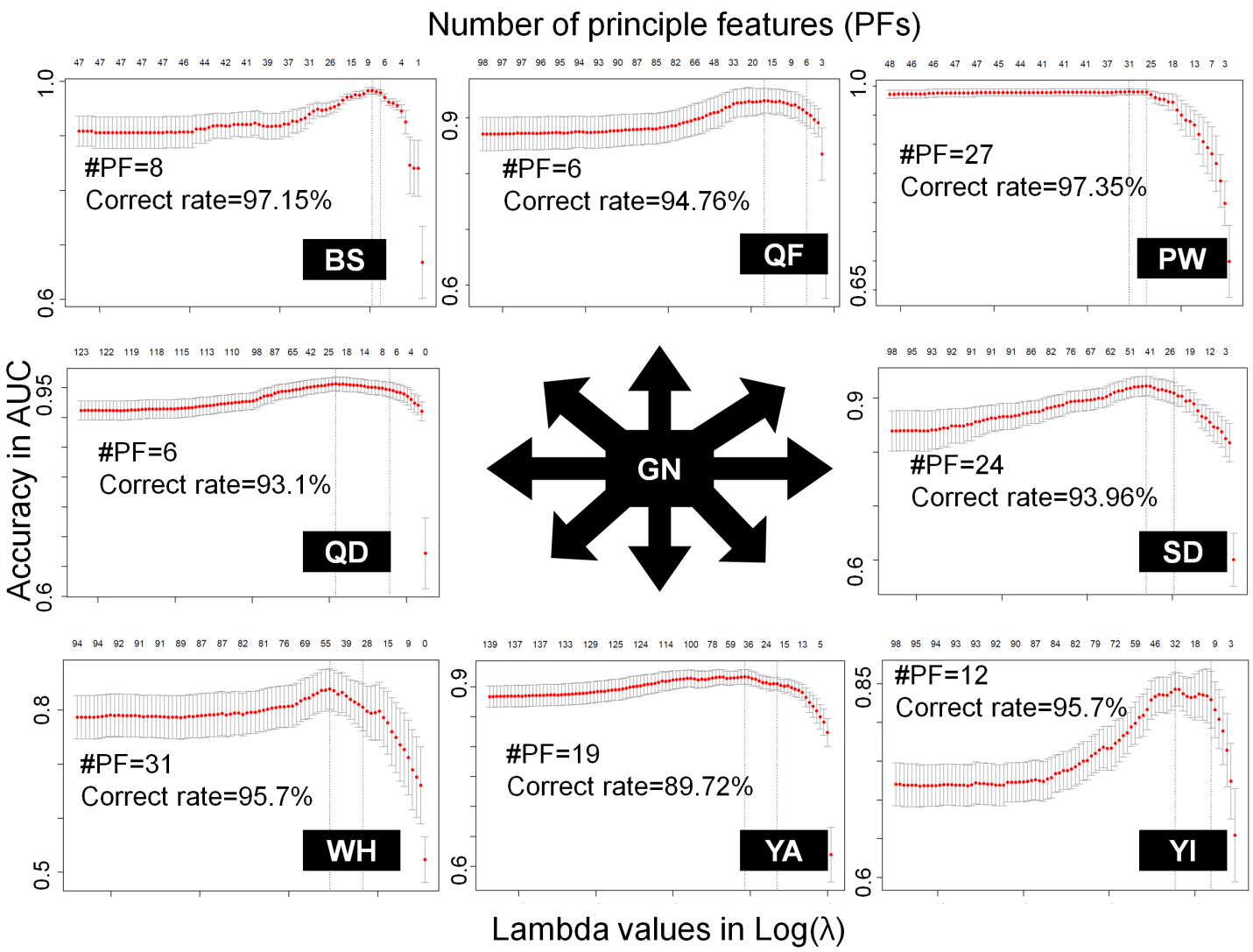

Figure 4.11: Binary classification performance for each BC along with its correct rate of prediction. GN constitution is set as common class for all eight pairs.

4.4 Pair-Wise Classification based on LASSO

By the proposed pair-wise classification on the basis of LASSO, the binary classification is conducted in stead of BC score regressions as three algorithms above. All the 851 samples are grouped into eight pairs according to the pre-process by CCMQ-J. Then, the eight pairs of binary classification are evaluated as shown in Fig. 4.11. The classification quality is evaluated by the accuracy of area under curve (AUC). For convenient observation, the correct rate of classification is also calculated as shown in each window. From this results, it is seen that the regularized average correct rate is 94.6% with 17 PFs for each BC in average. Compared to all methods above, the proposed pair-wise classification simplifies the task into distinguishing a specific biased BC against the GN constitution; and the distinction among eight biased BCs is not concerned. Therefore, the prediction and PF-extraction quality is higher but the pre-process is necessary.

In general, the proposed pair-wise classification based on LASSO achieves best accuracy and PF-extraction performances among all of the employed algorithms.

5 Cross-Algorithm Validation and Health-Guidance

In this chapter, the prediction results and identified principle features are summarized and compared. On the basis of such summary, the health-guidance of life-style is offered towards recovering the biased BCs into the "gentleness" state.

5.1 Cross-Algorithm Evaluation

The prediction performances of multiple algorithms are briefly summarized by Tab. 5.1. Among four algorithms of prediction, the pair-wise classification through LASSO achieves highest prediction accuracy 94.6% and fewest PFs 17 in average. As mentioned above, the goal of this work is to identify PFs for healthcare guidance. Therefore, the prediction quality is only considered as the confidence of obtained PFs. Specifically, we only concern the life-style items which distinguish a biased BC against the common class (GN). In this sense, the proposed pair-wise LASSO gives reliable hints that the non-principle features in life-style questionnaire are irrelevant for bias recovery.

Table 5.1: Comparisons among RF, PLS, LASSO and pair-wise classification over prediction quality and number of PFs

	RF	PLS	LASSO	pair-wise
Accuracy	88.7%	40.9%	69.9%	94.6%
Average PF #	254	28	31	17

The identified PFs addressing question IDs in our life-style questionnaire (the full list of question is seen in Tab. 3.2) are listed in Tab. 5.4. The total appearance

of PFs of PLS, LASSO and pair-wise LASS are 126, 84, and 77, respectively. However, some of PFs are duplicated in multiple BC types (with a maximum number of nine types) during the regression or classification. The PF appearance counting over BC types are illustrated in Fig. 5.1. For all of three algorithms, most PFs are relevant to the single or few BC predictions as expected. If a specific PF (seen as a life-style item) is involved in most of BC types, all nine types for instance, it indicates this item “critically" leads to nine BCs, which violates both of clinical and algorithmic fundamentals. In this sense, all of the PLS, LASSO and pair-wise classification give reasonable PF counting distributions.

Table 5.4: Cross algorithm validation for PFs

BC	QID of PFs by PLS	QID of PFs by LASSO	QID of PFs by pair-wise
	PLS	LASSO	pair-wise classification
GN	Q7, Q9, Q10, Q11, Q12, Q45, Q46, Q47, Q50, Q51, Q52, Q53, Q54, Q55, Q60, Q62, Q67, Q142, Q145, Q147, Q149, Q164, Q182, Q195, Q196, Q197, Q201, Q207, Q209, Q210, Q228, Q293, Q306, Q41, Q148, Q156, Q143	Q7, Q9, Q10, Q11, Q39, Q40, Q41, Q45, Q46, Q47, Q50, Q51, Q52, Q54, Q55, Q60, Q62, Q67, Q68, Q69, Q70, Q77, Q142, Q143, Q145, Q147, Q148, Q156, Q182, Q196, Q197, Q201, Q207, Q210, Q218, Q224, Q228	N/A

continued on next page

continued from previous page

BC	QID of PFs by PLS	QID of PFs by LASSO	QID of PFs by pair-wise
BS	Q148, Q155, Q156, Q159_1, Q187_1, Q217, Q218, Q230, Q278, Q289, Q304, Q320, Q57, Q58, Q59, Q60, Q61, Q62, Q63, Q141, Q142, Q147, Q153, Q199, Q201, Q202, Q204	Q45, Q47, Q52, Q53, Q55, Q57, Q58, Q59, Q60, Q61, Q62, Q63, Q66, Q141, Q142, Q147, Q153, Q156, Q183, Q187_25, Q187_28, Q196, Q198, Q199, Q201, Q202, Q203, Q204, Q207, Q210, Q217, Q218, Q230, Q302	Q51, Q53, Q56, Q57, Q61, Q203, Q210, Q230
PW	Q32, Q68, Q140, Q214, Q216, Q217, Q218, Q228, Q231, Q280, Q11, Q28, Q57, Q58, Q60, Q61, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q156, Q199, Q201, Q202, Q210, Q211, Q237	Q9, Q11, Q43, Q47, Q56, Q57, Q58, Q60, Q61, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q156, Q182, Q195, Q197, Q199, Q201, Q202, Q207, Q210, Q211, Q214, Q216, Q217, Q218, Q228, Q237, Q280	Q9, Q10, Q46, Q51, Q52, Q53, Q54, Q55, Q62, Q64, Q183, Q187_13, Q187_19, Q187_22, Q187_29, Q187_36, Q187_37, Q190, Q201, Q207, Q214, Q215, Q231, Q237, Q256, Q304, Q317
QD	Q41, Q148, Q218, Q10, Q44, Q49, Q50, Q51, Q54, Q56, Q142	Q7, Q8, Q10, Q11, Q12, Q41, Q42, Q43, Q44, Q47, Q49, Q50, Q51, Q52, Q54, Q55, Q56, Q57, Q59, Q60, Q62, Q63, Q64, Q66, Q67, Q69, Q141, Q142, Q148, Q182, Q195, Q198, Q202, Q203, Q210, Q218, Q234	Q8, Q10, Q47, Q51, Q52, Q54

continued on next page

continued from previous page

BC	QID of PFs by PLS	QID of PFs by LASSO	QID of PFs by pair-wise
QF	Q31, Q40, Q143, Q187_25, Q187_42, Q217, Q218, Q224, Q227, Q317, Q333, Q47, Q57, Q59, Q60, Q61, Q62, Q63, Q64, Q65, Q66, Q145, Q182, Q183, Q198, Q228	Q7, Q8, Q9, Q11, Q12, Q40, Q46, Q47, Q50, Q51, Q52, Q54, Q55, Q57, Q60, Q62, Q63, Q64, Q65, Q66, Q67, Q69, Q70, Q142, Q143, Q145, Q147, Q182, Q187_5, Q195, Q198, Q199, Q201, Q202, Q203, Q210, Q218, Q224, Q228, Q316	Q8, Q10, Q47, Q54, Q63, Q210
SD	Q35, Q37, Q38, Q159_3, Q185, Q187_1, Q187_34, Q217, Q218, Q242, Q311, Q312, Q319, Q57, Q59, Q60, Q62, Q63, Q64, Q66, Q142, Q187_15, Q187_16, Q187_18, Q187_24, Q198, Q199, Q201	Q7, Q9, Q44, Q47, Q57, Q59, Q60, Q62, Q63, Q64, Q66, Q67, Q142, Q183, Q187_16, 187_24, Q195, Q199, Q201, Q207, Q217, Q218	Q9, Q11, Q12, Q40, Q47, Q51, Q54, Q55, Q59, Q60, Q62, Q66, Q74, Q149, Q187_5, Q187_16, Q187_18, Q187_24, Q187_28, Q195, Q207, Q218, Q270, Q324
WH	Q32, Q33, Q37, Q70, Q76, Q77, Q78, Q187_12, Q187_19, Q214, Q217, Q218, Q228, Q247, Q252, Q267, Q282, Q290, Q319, Q43, Q44, Q60, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q201, Q202, Q219, Q224, Q237, Q306	Q9, Q11, Q33, Q43, Q44, Q47, Q53, Q56, Q59, Q60, Q62, Q63, Q64, Q66, Q141, Q142, Q147, Q195, Q196, Q201, Q202, Q207, Q217, Q218, Q237, Q306	Q5, Q7, Q8, Q9, Q11, Q43, Q44, Q51, Q52, Q53, Q57, Q60, Q63, Q67, Q140, Q141, Q142, Q149, Q187_30, Q210, Q218, Q262, Q273, Q275, Q279, Q295, Q299, Q300, Q311, Q317, Q333

continued on next page

continued from previous page

BC	QID of PFs by PLS	QID of PFs by LASSO	QID of PFs by pair-wise
YA	Q8, Q43, Q155, Q156, Q187_1, Q187_11, Q216, Q218, Q224, Q227, Q232, Q253, Q257, Q258, Q259, Q309, Q316, Q322, Q57, Q59, Q60, Q61, Q62, Q63, Q64, Q66, Q142, Q153, Q187_16, Q187_9, Q198, Q199, Q201, Q202, Q228, Q306	Q7, Q11, Q47, Q57, Q59, Q60, Q61, Q62, Q63, Q64, Q74, Q77, Q142, Q153, Q156, Q182, Q187_9, Q187_15, Q187_16, Q187_23, Q196, Q198, Q201, Q202, Q216, Q218, Q227, Q306	Q11, Q47, Q51, Q54, Q57, Q59, Q60, Q62, Q64, Q68, Q74, Q77, Q142, Q156, Q182, Q183, Q201, Q210, Q218
YI	Q32, Q78, Q155, Q156, Q160, Q187_33, Q217, Q218, Q252, Q253, Q328, Q57, Q59, Q61, Q62, Q63, Q66, Q142, Q199, Q64, Q60, Q153, Q201, Q202	Q7, Q9, Q45, Q47, Q57, Q59, Q60, Q62, Q63, Q65, Q66, Q67, Q142, Q153, Q201, Q202, Q207, Q218	Q10, Q12, Q45, Q47, Q54, Q59, Q62, Q64, Q66, Q159_3, Q187_34, Q201

5.2 Health-Guidance

It is also expected to reduce the number key items from the clinical point of view since the healthcare-guidance over dozens of life-styles is still unpractical. Moreover, a single algorithm can hardly offer convincing PFs due to the indeterminacy of machine learning and data-set itself. Thus, the cross-algorithm guidance is necessary. The straightforward manner is to summarize the common PFs for each BC obtained by all algorithms as follows.

5.2.1 For gentleness

You are well, please keep it.

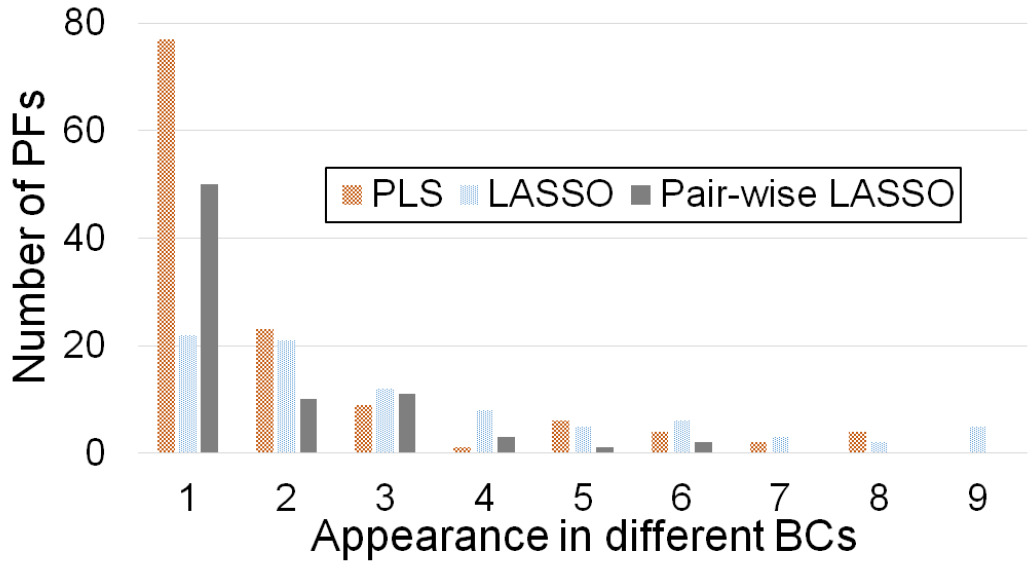

Figure 5.1: Counting of identified PFs appear in multiple BC predictions for PLS, LASSO and pair-wise classification algorithms

5.2.2 For blood-stasis

Two common PFs for BS: (1) Q57 (Job Stress): conditions for the past month: I feel dizzy. (2) Q61 (Job Stress): conditions for the past month: I have a backache.

Guidance for recovery (from BS to GN): since two items related to the job stress are identified by cross-algorithms, the persons who have been heavily involved in office works are in high risk of falling in BS. For recovery to GN, they are expected to (1) avoid long-time on-sit works by breaking them with body-excises in very one hour at least; (2) keep well maintains of office ventilation; (3) adjust proper posture during the desk-works.

5.2.3 For Phlegm-wetness

Five common PFs for PW: (1) Q62 (Job Stress): conditions for the past month: I have eyestrain. (2) Q64 (Job Stress): conditions for the past month: My stomach is not in good shape. (3) Q201 (Sleep State): you had trouble sleep in because you feel too cold. (4) Q214 (Sub-Health): About how you chew, please

choose one applicable to you. (5) Q237 (Dietary Habits): How much Chinese tea (Oolong, Jasmine, Pu're tea) do you drink per time?

Guidance for recovery (from PW to GN): (1) improve the sleep time and sleep environment, and take a multivitamins (A, C, and E, especially); (2) chew slowly, increase and reduce fiber and fat intakes; (3) try to drink (reasonably more) tea since tea polyphenols help improving liver function and diuretic.

5.2.4 For Qi-depression

Three common PFs for QD: (1) Q10 (Depressed mood): depressed that nothing could cheer you up. (2) Q51 (Job Stress): conditions for the past month: Depressed. (3) Q54 (Job Stress): conditions for the past month: I am depressed.

Guidance for recovery (from QD to GN): The most critical and common factor of QD is identified as the mental situation (depression in particular). Therefore, the mental management and psychological consulting are considered as most effective manner for PW recovery. Specifically and softly, the persons with PW are suggested to activate in social connections such as joining social events as the early efforts. The further psychological interventions are quite sophisticated and out of scope of this article.

5.2.5 For Qi-deficiency

Two common PFs for QF: (1) Q47 (Job Stress): conditions for the past month: Dull. (2) Q63 (Job Stress): conditions for the past month: I feel my heart pounding or I get out of breath.

Guidance for recovery (from QF to GN): Similarly to QD, QF relevant persons appear mental un-balance (high tense in particular). Therefore, relaxes such as Yoga, well sleep, vacations, entertainments, even deep-breath are suggested.

5.2.6 For special diathesis

Seven common PFs for SD: (1) Q59 (Job Stress): conditions for the past month: I feel heavy in the head or have a headache. (2) Q60 (Job Stress):

conditions for the past month: I have bad stiff neck and shoulders. (3) Q62 (Job Stress): conditions for the past month: I have eyestrain. (4) Q66 (Job Stress): conditions for the past month: I suffer from diarrhea or constipation. (5) Q187_16 (Disease): Allergic rhinitis. (6) Q187_24 (Disease): Atopic dermatitis. (7) Q218 (Dietary Habits): Do you feel dry in your mouth?

Guidance for recovery (from SD to GN): (1) avoid long-time sitting, adjust proper posture for desk-works, and warm bath; (2) increase intakes of multivitamin (A, C, and E especially); (3) avoid the stimulus diet since the persons related to SD appear irritable bowel syndrome (IBS); (4) try to identify some allergic origin on foods since they appear allergic inflammation somehow; (6) hydration.

5.2.7 For Wet-heat

Six common PFs for WH: (1) Q43 (Job Stress): conditions for the past month: Irritated. (2) Q44 (Job Stress): conditions for the past month: Annoyed. (3) Q63 (Job Stress): conditions for the past month: I feel my heart pounding or I get out of breath. (4) Q141 (Personality): self-assessment: I see myself as critical, quarrelsome. (5) Q142 (Personality): self-assessment: I see myself as anxious, easily upset. (6) Q218 (Dietary Habits): Do you feel dry in your mouth?

Guidance for recovery (from WH to GN): relaxes such as Yoga, well sleep, vacations, entertainments, even deep-breath are suggested. (2) hydration without increment of sugar intakes (namely, drink more pure water instead of juice).

5.2.8 For Yang-deficiency

Nine common PFs for YA: (1) Q57 (Job Stress): conditions for the past month: I feel dizzy. (2) Q59 (Job Stress): conditions for the past month: I feel heavy in the head or have a headache. (3) Q60 (Job Stress): conditions for the past month: I have bad stiff neck and shoulders. (4) Q62 (Job Stress): conditions for the past month: I have eyestrain. (5) Q64 (Job Stress): conditions for the past month: My stomach is not in good shape. (6) Q142 (Personality): self-assessment: I see myself as anxious, easily upset. (7) Q156 (Body Index):

Weight. (8) Q201 (Sleep State): you had trouble sleep in because you feel too cold. (9) Q218 (Dietary Habits): Do you feel dry in your mouth?

Guidance for recovery (from YA to GN): the PFs of YA are complicated in contrast to other BCs. It is suggested to improve the mental and diet status similarly to all above. Moreover, the hydration along with sugar intakes is suggested.

5.2.9 For Yin-deficiency

Four common PFs for YI: (1) Q59 (Job Stress): conditions for the past month: I feel heavy in the head or have a headache. (2) Q62 (Job Stress): conditions for the past month: I have eyestrain. (3) Q66 (Job Stress): conditions for the past month: I suffer from diarrhea or constipation. (4) Q201 (Sleep State): you had trouble sleep in because you feel too cold.

Guidance for recovery (from YI to GN): (1) mental management; (2) intakes of multivitamin; (3) avoid the stimulus diet since the persons related to SD appear irritable bowel syndrome (IBS).

5.3 Individual Recovery Perspective

The sections above offer the general health-care guidance for each BC. Namely, the specific situation of each individual is not distinguished but the overview-based summary is made. From the pair-wise LASSO results, it is also possible to offer the recovery guideline for specific individuals. In this case, only the identified PFs of each pair are concerned. The biased BCs are not fair for each individual. The bias levels and properties vary among all of tested persons. In other words, it is necessary to analyze "how bias he/she is" and "which life-styles should be efficiently changed" in detail. Thus, the concept of "bias-degree" is defined in this thesis. As shown in Fig. 5.2, (for a specific pair of GN-to-bias) all the samples of GN are projected into a highly dimensional space, where the dimensions are referred to the identified PFs. The center of GN cluster (seen as gravity) can be calculated by averaging GN samples in terms of Euclidean distance for instance. Then, the Euclidean distance from a specific sample to the center is defined as D_i. The hyperspheres can be found around the gravity with the radial

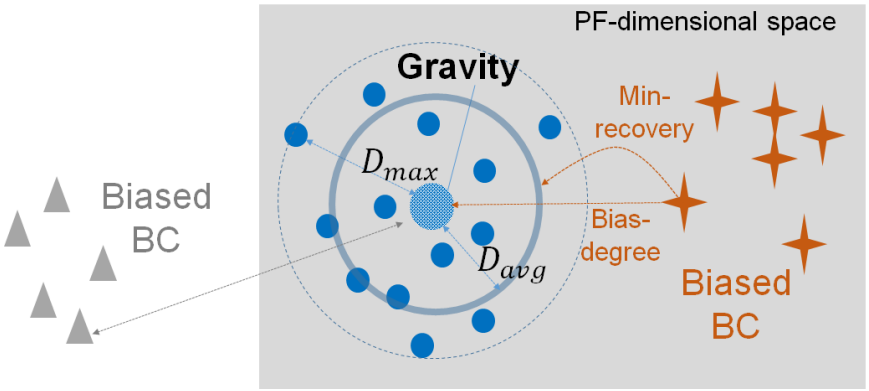

Figure 5.2: Pair-wise LASSO results are used to calculate bias-degree and minimum efforts for recovering to GN.

volumes D_{max} and D_{avg} which describe the furthest and average distance to the GN center. Considering a biased BC, each sample appears different distance to the GN center D_{bias_i} seen as the bias-degree. To recover towards GN, the biased sample is expected to shift in to the radial D_{avg}. The difference Δ between the bias-degree and D_{avg} indicates the minimum effort for this individual recovering to GN. The minimum efforts can be given in the following equation:

$$\Delta = \sqrt{{D_{bias\_i}}^2 - {D_{avg}}^2}. \tag{5.1}$$

Table 5.5 illustrates one example of individual recovery guideline referring the samples from the GN-BS pairing. In this example, eight PFs are identified by the proposed pair-wise LASSO algorithm (Q51, Q53, Q56, Q57, Q61, Q203, Q210, Q230); and 15 persons from our data-base are diagnosed as BS. In such an eight-dimensional space, the center of GN is given as the referee for the recovery. Namely, the specific individual is suggested to change his/her life-style towards these eight indexes. The column "min recover" indicates the gap requirement for him/her to return to GN as a total score. Obviously, how to fill this gap depends on the case-by-case consulting. For instance, the individual with ID number of 60 is diagnosed as BS by CCMQ-J; and the gap for him/her to recover towards GN is 3.75. This individual is expected to change the eight life-styles orienting the referee with the total score of 3.75. It is consultable to assign this 3.75 points to feasible items from his/her personal situation.

Table 5.5: Example of recovering BS to GN with 8 PFs for all 15 individuals

FP ID	Q51	Q53	Q56	Q57	Q61	Q203	Q210	Q230	D_{avg}
GN-referee	1.76	1.68	1.55	1.24	1.81	1.29	1.57	4.04	1.91
ID									Min recover
60	3	4	2	1	4	3	3	3	3.75
148	1	1	1	1	4	3	2	4	2.31
189	3	1	3	1	3	4	3	4	3.32
207	2	2	2	4	4	2	3	3	3.52
268	3	4	3	1	2	3	3	4	3.18
283	4	3	3	2	3	2	3	3	3.24
340	3	4	1	2	2	4	4	3	4.26
487	4	4	3	2	4	1	3	4	4.00
574	4	3	3	2	1	2	2	2	3.30
649	2	2	3	4	3	2	1	2	3.50
658	3	2	3	3	3	3	3	3	3.21
683	3	3	3	2	1	4	3	3	3.62
750	4	4	2	2	3	2	3	3	3.49
779	3	2	4	2	3	2	2	3	2.72
830	2	2	2	2	4	2	3	3	2.32

All of eight pairs of biased BCs against GN along with the corresponding GN gravity and average distances to gravity are listed in Tab. 5.6. A specific individual can be evaluated by the relevant PF items of his/her BC type. The Euclidean distance between this evaluation vector and the corresponding gravity is calculated and compared with the average distance D_{avg}. It is expected to compensate the gap towards the GN similarly to the above example. This table is seen as the "Referee Table" as shown in the implementation flow. For the personal consulting of BC treatment, the individual is only expected to answer the questions appearing in this table according to his/her BC types. Then, the clinic doctor will refer to the "Referee Table" and calculate the minimum efforts for him/her to recover to GN. At last, a personal plan will be suggested to fill the scoring gap.

Table 5.6: Eight pairs of biased BCs against GN along with corresponding GN gravity and average distances to gravity

	gravity of principle features	D_{avg}
BS-GN	Q51=1.76 , Q53=1.68 , Q56=1.55 , Q57=1.24 , Q61=1.81 , Q203=1.29 , Q210=1.57 , Q230=4.04 ,	1.91
PW-GN	Q9=0.96 , Q10=0.97 , Q46=1.78 , Q51=1.76 , Q52=1.81 , Q53=1.68 , Q54=1.79 , Q55=1.42 , Q62=2.28 , Q64=1.51 , Q183=1.15 , Q187_13=0.00 , Q187_19=0.02 , Q187_22=0.04 , Q187_29=0.08 , Q187_36=0.12 , Q187_37=0.12 , Q190=7.20 , Q201=1.33 , Q207=2.07 , Q214=3.90 , Q215=1.47 , Q231=3.74 , Q237=2.58 , Q256=1.87 , Q304=0.72 , Q305=0.49 , Q317=3.06 ,	5.42
QD-GN	Q8=0.59 , Q10=0.97 , Q47=1.97 , Q51=1.76 , Q52=1.81 , Q54=1.79 ,	1.72
QF-GN	Q8=0.59 , Q10=0.97 , Q47=1.97 , Q54=1.79 , Q63=1.25 , Q210=1.57 ,	1.68
SD-GN	Q9=0.96 , Q11=1.14 , Q12=0.84 , Q40=2.38 , Q47=1.97 , Q51=1.76 , Q54=1.79 , Q55=1.42 , Q59=1.59 , Q60=2.19 , Q62=2.28 , Q66=1.71 , Q74=2.81 , Q149=4.25 , Q187_5=0.11 , Q187_16=0.20 , Q187_18=0.06 , Q187_24=0.08 , Q187_28=0.02 , Q195=2.03 , Q207=2.07 , Q218=1.83 , Q270=3.27 , Q324=1.83 ,	4.47
WH-GN	Q5=0.00 , Q7=0.81 , Q8=0.59 , Q9=0.96 , Q11=1.14 , Q43=1.89 , Q44=1.98 , Q51=1.76 , Q52=1.81 , Q53=1.68 , Q57=1.24 , Q60=2.19 , Q63=1.25 , Q67=1.59 , Q140=3.40 , Q141=2.73 , Q142=4.03 , Q149=4.25 , Q187_30=0.12 , Q210=1.57 , Q218=1.83 , Q262=1.67 , Q273=1.31 , Q275=0.74 , Q279=1.23 , Q295=1.43 , Q299=0.97 , Q311=0.95 , Q317=3.06 , Q333=1.56 ,	5.29
YA-GN	Q11=1.14 , Q47=1.97 , Q51=1.76 , Q54=1.79 , Q57=1.24 , Q59=1.59 , Q60=2.19 , Q62=2.28 , Q64=1.51 , Q68=2.61 , Q74=2.81 , Q77=2.31 , Q142=4.03 , Q156=60.65 , Q182=2.05 , Q183=1.15 , Q201=1.33 , Q210=1.57 , Q218=1.83 ,	11.37
YI-GN	Q10=0.97 , Q12=0.84 , Q45=2.11 , Q47=1.97 , Q54=1.79 , Q59=1.59 , Q62=2.28 , Q64=1.51 , Q66=1.71 , Q159_3=0.05 , Q187_34=0.00 , Q201=1.33 ,	2.02

5.4 Discussion

In this thesis, multiple ML algorithms are applied to predict the BCs from the very highly dimensional samples. For some of other engineering applications such as computer vision or audio, a specific but powerful algorithm may fit the specific task well. However, the prediction accuracy is not only one of (even out of) scopes of consideration. As shown in Fig. 5.3, our strategy is to identify the common characteristics from various ML efforts. In this sense, a general conclusion of this thesis is beyond the BC prediction itself but the ubiquitous philosophy of medical

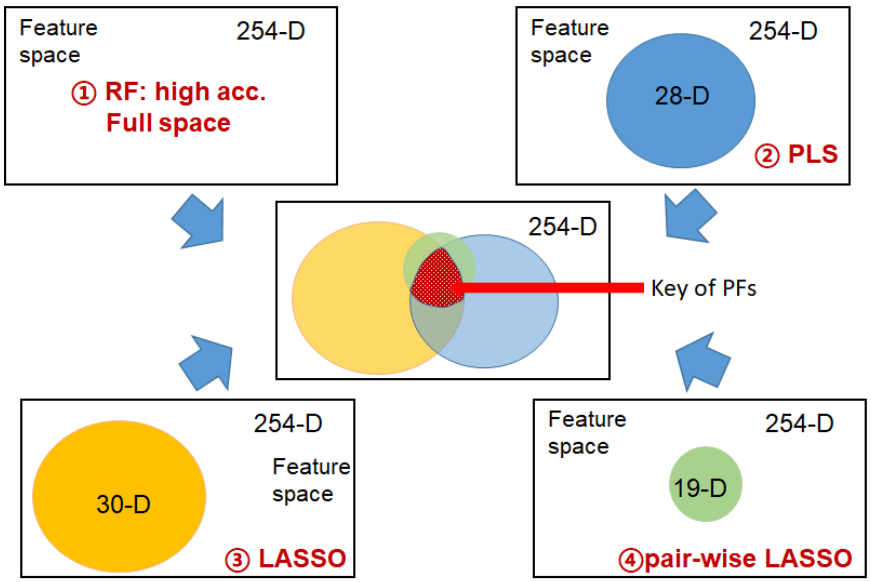

Figure 5.3: Strategy of cross-algorithm validation for PF identification

data analysis as: various algorithms cross-validation is more trustable than any single one.

The validation of proposed (in fact, any efforts of) methods of prediction and analysis should be based on the well-developed, highly trustable, sufficiently big, and efficient data-base. In this project, only 851 persons are involved in our experiment. As a start point of scientific data analysis of BC theory, our original data-base appears well performances. However, the amount of 851 is far smaller than other machine learning data-base even the traditional medical data-base. Therefore, most of inaccuracy and PF defects are supposed to come from the poor scale of our data-base. Along with our data-base becomes bigger (hopefully), the analysis strategy to deal with "big data" will be more noticeable. Here, the methods to "reduce dimensions" are illustrated; the appendix B illustrates the methods to "reduce samples", which may not be necessary in current small data-base but will be critical if the amount of individuals increases to million.

6 Conclusion

The body constitutions (BCs) of traditional Chinese medical theory are predicted through machine learning algorithms in this work. On the basis of the original questionnaire including 254 life-style features, the ML algorithms including RF, PLS, LASSO and a new scheme of pair-wise classification are employed for predicting the BCs over the population of 851 persons. Moreover, the principle features (PFs) of life-style are identified to recover the biased BCs into the gentle constitutions as the health guidance.

From the prediction results, a maximum correction rate of 88.7%, 40.9% , 69.9%, 94.6% are achieved by RF, PLS, LASSO, and our proposed pair-wise classification algorithm, which indicates the developed life-style questionnaire is effective. Meanwhile, the principle features identified by above algorithm are extracted with an average number of 17. The total appearance of PFs of PLS, LASSO and pair-wise LASS are 126, 84, and 77, respectively. By a cross-algorithm validation, most commonly identified principle life-styles are illustrated for each biased BC. Furthermore, the health care guidance is suggested for the bias recovery to the gentle constitution. As the conclusion, the ML algorithms are trustable and fair for explicit applications, but sensitive to implicit applications such as life-style or body constitutions. Differed algorithms do not lead to incorrect clinical explanations but the quality of application might be quite different. Then, it is necessary to carefully choose algorithms (even multiple validations) for analyzing implicit medical data; and the pre-process before any MLs is suggested.

The statistical results from our proposed method are correctly explainable for the nine types of body constitutions in multifaceted viewpoints of health conditions and understandable through existing medical theory; moreover, new medical hints, which have not been explicitly indicated by traditional medical researches, have been extracted from our data analysis. Crowdsourcing makes it possible to

accumulate data from the large network of potential participants and bring people together to harness their collective information. Therefore, it is an efficient methodology to verify an important scientific assumption.

7 Acknowledgement

I would like to express my sincere gratitude towards the following people. The Doctotaral degree and the completion of the thesis would become impossible without their support and encouragement.

First of all, I would like to sincerely express my thanks to my supervisor, Professor Shigehiko Kanaya, for his extremely valuable guidance and powerful supports during the past three years. His enthusiasm in teaching and research and his enduring encouragement led me to become a full fledged person. It is my great honor and fortunate to have taken him as my supervisor, and the experiences in this laboratory will be the treasure in my whole life. I would like to thank Professor Keiichi Yasumoto, Associate Professor MD.Altaf-Ul-Amin, Associate Professor Naoaki Ono, and assistant Professor Alex Ming Huang, for their preview of the thesis and their extremely valuable comments to my research. Their remarkable suggestions were indispensable for making my dissertation study successful.

I express my appreciation to Ms. Minako Ohashi for her help on so many documentaries, and to Ms. Aki Morita for her helpful support. I wish to express my sincere appreciation to those who have helped me in the past three years. I want to express my gratitude to all the lab members, especially Mr. Zheng Chen. During these three years, I received the most encouragement from them. I am so lucky to study and work with all of them.

I would like to thank my family. Whenever I need help, they are always there. Their support is of great importance and really precious to me. Without their warm concerns and helps, it would be impossible for me to complete my study. I feel their warm care and support, which encourages me all the time. I also appreciate all my friends. Being with them made my college life an amazing experience.

Appendix A: Abbreviations

TCM: traditional Chinese medicine;
BC: body constitution;
ML: machine learning;
GN: gentleness;
QF: Qi-deficiency;
YA: Yang-deficiency;
YI: Yin-deficiency;
PW: Phlegm-wetness;
WH: Wet heat;
BS: Blood-stasis;
QD: Qi-depression;
SD: Special diathesis;
RF: random forest;
PLS: partial least squares;
LASSO: least absolute shrinkage and selection operator;
PF: principle feature;
AUC: area under curve.

Appendix B: Sample Reduction of Medical Database

As mentioned in the main context, both of dimension reduction and sample reduction are important for medical data analysis. Our original BC-relevant database is very initial trial of this research field. So far, only 851 samples are available for both of CCMQ-J and life-style questionnaire. Along with the project going on, the number of samples will increase to large. Till then, the machine learning algorithms for reducing the sample space and identifying the principle samples are necessary. The following contents demonstrate how the support vector machine (SVM) algorithm effects medical data analysis by reducing the samples.

SVM in Medical Science

This appendix shows the efficient and practical scheme of medical data analysis through machine learning algorithms. The support vector machine (SVM) mechanism is specifically employed for building an artificial intelligence (AI) assistant diagnosis systems. Considering the practical demands on clinical diagnosis, the naive SVM algorithm is hardly used since the poor number of classes (typically, two classes) and explosion of samples. Therefore, a sample domain description technology is developed to realize a one-class SVM for flexibly expending the number of classes. Furthermore, an on-line learning strategy is proposed to implement high-performance classification/diagnosis with greatly reduced database. For proof-of-concept, several medical databases are employed for diagnosis test. From the test results, the diagnosis correct-rate is improved with compact database; and the scale of database is reduced while the similar correct-rate is achieved by naive SVM algorit

Since most of diagnosis systems are built through data classification or so-called pattern recognition networks, this thesis focuses on an advanced classification algorithm support vector machine (SVM) in particular. From some laboratory works in medical science, SVM performs very high correct rate in categorizing medical cases which is described by the high-dimensional feature vectors. Employing a sufficiently large database of disease cases with correct labels, SVM classifiers are built by training process. Then, the well-trained classifier accepts new cases of which the category labels are unknown ("ill" or "healthy" for instances). The classifier, known as diagnosis system, gives the predicted labels instead of clinical doctors' judge.

Several real-world medical databases are employed for proof-of-concept. After constructing the SVM classifier, the most significant disease cases (known as support vectors (SVs)) are identified and offered to clinical doctors. The proposed system can diagnose the new case with very high accuracy. More importantly, the SVs are offered and referenced by doctors for assisting the judgment. Obviously, the scale of database and the number of SVs are usually very large through the conventional SVM implementations. Thus, **the essence of this appendix is making the most efficient use of the database and minimize the number of SVs for practical applications.** For this purpose, a novel on-line learning SVM scheme is proposed and verified by medical database tests. Furthermore, the novel data domain description algorithm is developed for flexible number of classes, which is usually two-class classification by conventional SVM. From the database test results (including breast-cancer, heart disease and liver disorder), the number of SVs is greatly reduced with a high accuracy on diagnosis; and a single class SVM is achieved with very few SVs.

Similarly to many reported works, the SVM algorithm can be applied to classify a highly dimensional vector $\mathbb{X}$s with the form of $\mathbb{X} = (x_1, x_2, \ldots, x_n)$. When this vector represents the feature of a medical case, the classification label can be considered as an AI diagnosis result. The process to obtain a suitable math-model for the classification is called "SVM training process", which is out of this thesis summary. The principle of SVM training and classification is shown in Fig. 7.1 along with an example of breast-cancer database. In this example, the 689 cases of real breast-cancer features are introduced with ten dimensions, and labeled

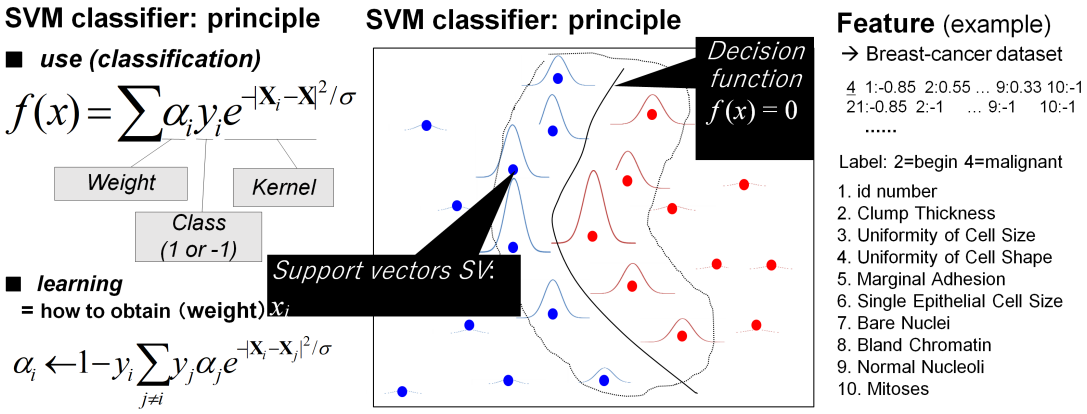

Figure 7.1: Medical data is classified by SVM along with an example of breast-cancer dataset.

by "2" as begin and "4" as malignant. The conventional SVM algorithm with Gaussian kernels is employed to pursue the prediction model from the database. The prediction function is given by:

$$f(\mathbb{X}) = sign[\sum_{i=1}^{N} \alpha_i y_i K(\mathbb{X}, \mathbb{X}_i) + b], \tag{7.1}$$

where the $\mathbb{X}_i$ is the support vectors, which are selected from the database through SVM training:

$$\alpha_i \leftarrow 1 - y_i(\sum_{j(\neq i)} \alpha_j y_j K(\mathbb{X}_i, \mathbb{X}_j) + b). \tag{7.2}$$

The conventional SVM is applied to diagnose the breast-cancer (see Fig. 7.2). When the number of support vectors (SVs) is close to entire database, the accuracy is perfect; however, when the number of SVs is smaller, the accuracy is going very poor. As mentioned above, the goal of AI diagnosis assistant system is not AI diagnosing but offering SVs (known as important reference cases) to doctors. A large number of SVs is impossible for doctors to review on site for any specific patient. Thus, conventional SVM can not be directly applied in clinic even though many works claimed that very high accuracy can be achieved by a huge database and complex math-model.

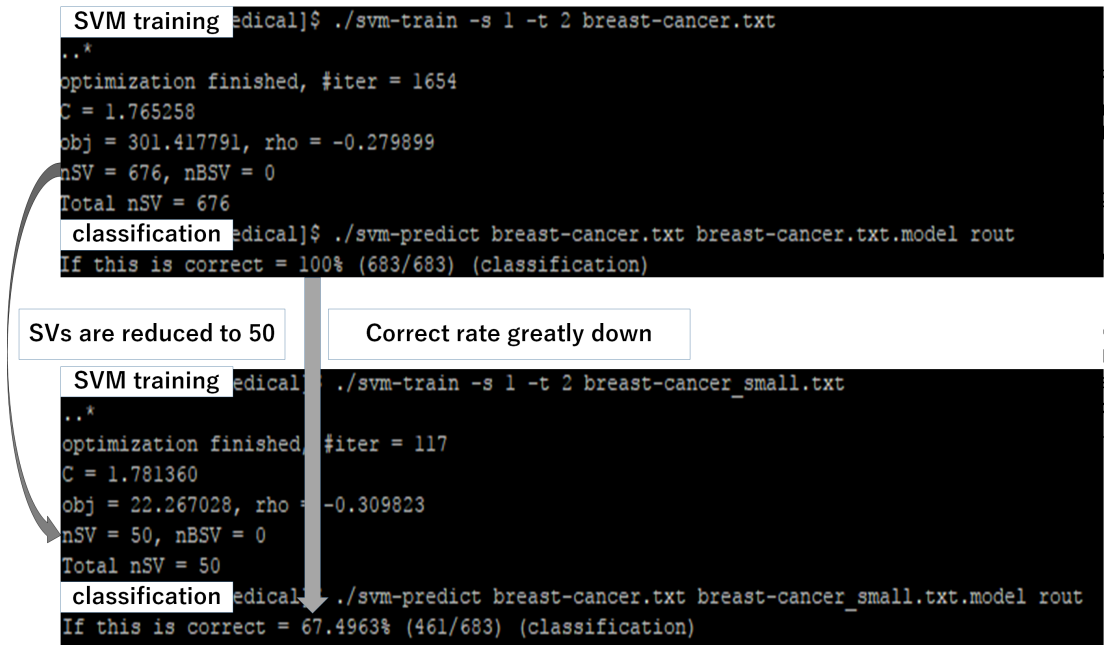

Figure 7.2: Conventional SVM is applied for breast-cancer diagnosis. When number support vectors is reduced, accuracy is going poor.

On-line Learning SVM for Diagnosis Assistant System

In the real-world applications SVMs, the number of learning samples is usually very large and unpredictable. As a result, large database or traditional on-line incrementally learning strategies can be hardly implemented in the clinical diagnosis assistant systems. Fortunately, in SVM theory some of the learning samples (non-support vectors) are ineffective, which can be removed from sample space. A "doctor-friendly" on-line learning strategy with constant number of learning samples is proposed in this work. In order to reduce the loss of accuracy, the effectiveness of each sample is evaluated and only the most ineffective sample is replaced by an on-line pattern. In this manner, the learning sample space can be expanded within compact SV space (see Fig. 7.3). The process of this on-line learning strategy is shown as follows:

1. Initial SVM learning according to a small set of samples;

2. Classifying the new-received on-line pattern;

3. Evaluating the effectiveness of previous samples and replacing the most inefficient one by new-received pattern;

4. SVM learning according to the updated samples;

5. Receiving new on-line pattern and repeating 2, 3, and 4.

After sufficient on-line learning operations, all the inefficient samples are replaced by significant on-line patterns. Then, the small scale of SVs are offered to the doctor for reviewing and referencing. This SV space is dynamically updated along with the career of specific doctors.

From the right part of Fig. 7.3, it is obviously found the proposed decremental on-line SVM achieves higher correct rate when the same (reduced) scale of database is applied by conventional SVM. Namely, the minimum necessary number of SVs is reduced with the same consideration of correct rate. Making breast-cancer as example: the proposed on-line SVM achieves perfect correct rate when the number of SVs is about 80; for the same correct rate, the conventional SVM needs at least 360 SVs. Other experiments reflect similar property of the proposed on-line learning process.

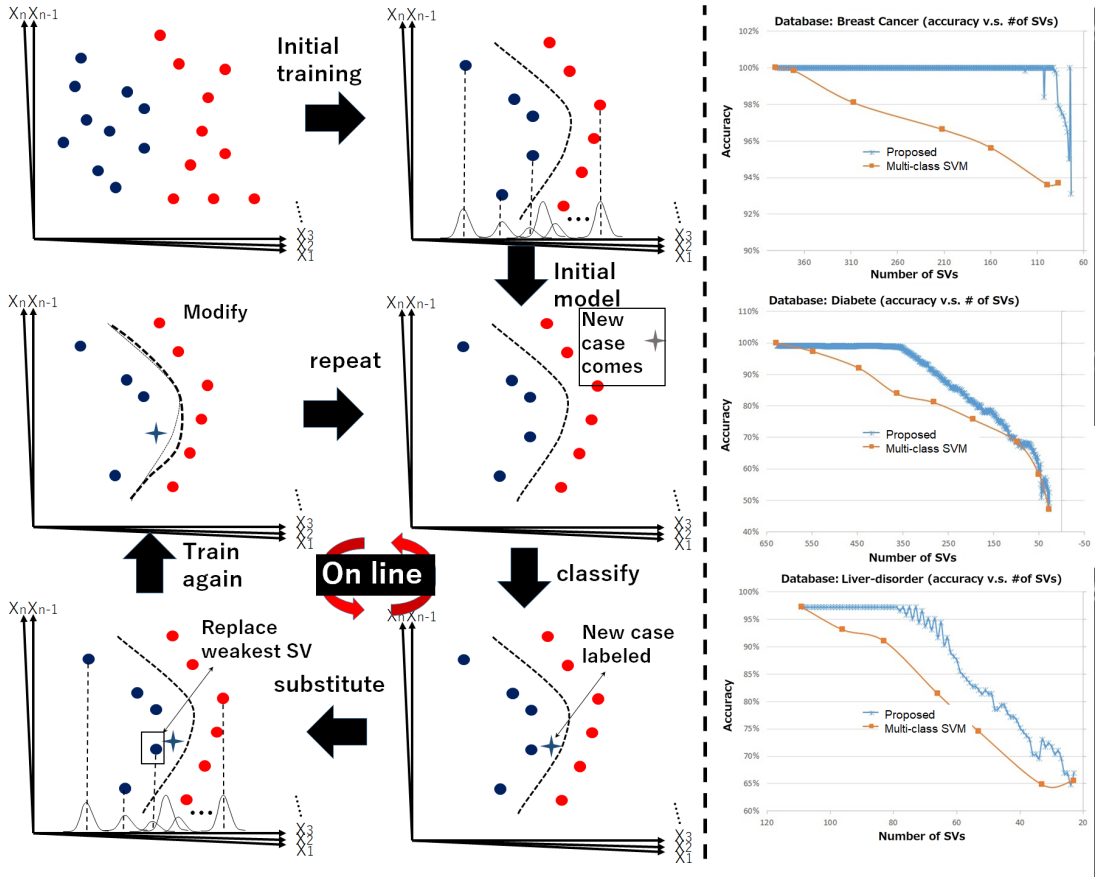

Figure 7.3: Principle of proposed on-line learning SVM strategy along with three examples: breast-cancer, heart disease, and liver disorder database

References

[1] Comprehensive survey of living conditions. *https://www.mhlw.go.jp/english/database/db-hss/cslc-index.html*.

[2] Konosuke Asanou, Naoaki Ono, Chika Iwamoto, Kenoki Ohuchida, Koji Shindo, and Shigehiko Kanayak. Feature extraction and cluster analysis of pancreatic pathological image based on unsupervised convolutional neural network. pages 2738–2740, 12 2018. doi: 10.1109/BIBM.2018.8621323.

[3] Qian Bai, Yaochen Chuang, Yonghua Zhao, Yao Wang, Pu Ge, Youhua Xu, and Ying Bian. The correlation between demographical and lifestyle factors and traditional chinese medicine constitution among macau elderly individuals. *Evidence-Based Complementary and Alternative Medicine*, 2021: 1–9, 04 2021. doi: 10.1155/2021/5595235.

[4] Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakhan, Busarakum Chaitusaney, Nattapong Jaimchariyatam, Ekapol Chuangsuwanich, Wei Chen, Huy Phan, Nat Dilokthanakul, and Theerawit Wilaiprasitporn. Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning(provide datasets). *IEEE Journal of Biomedical and Health Informatics*, PP, 11 2020. doi: 10.1109/JBHI.2020.3037693.

[5] András Benedek, György Molnár, and Szűts Zoltán. Practices of crowdsourcing in relation to big data analysis and education methods. 09 2015. doi: 10.1109/SISY.2015.7325373.

[6] Johannes Bircher. Towards a dynamic definition of health and disease. *Medicine, health care, and philosophy*, 8:335–41, 02 2005. doi: 10.1007/s11019-005-0538-y.

[7] Rose Chan. Concepts of body constitution, health and sub-health from traditional chinese medicine perspective. *World Journal of Translational Medicine*, 2:56, 01 2013. doi: 10.5528/wjtm.v2.i3.56.

[8] Shih-Lin Chen, Yun-Ting Liu, Kuang-Chieh Hsueh, and Pei-Ling Tang. Body constitution of traditional chinese medicine caused a significant effect on depression in adult women. *Complementary Therapies in Clinical Practice*, 42:101288, 2021. ISSN 1744-3881. doi: https://doi.org/10.1016/j.ctcp.2020.101288. URL `https://www.sciencedirect.com/science/article/pii/S1744388120311634`.

[9] Zheng Chen, Naoaki Ono, Wei Chen, Toshiyo Tamura, MD Altaf-Ul-Amin, Shigehiko Kanaya, and Ming Huang. The feasibility of predicting impending malignant ventricular arrhythmias by using nonlinear features of short heartbeat intervals. *Computer Methods and Programs in Biomedicine*, 205:106102, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106102.

[10] Alarcos Cieza, Cornelia Oberhauser, Jerome Bickenbach, Somnath Chatterji, and Prof. Dr. med. Gerold Stucki. Towards a minimal generic set of domains of functioning and health. *BMC public health*, 14:218, 03 2014. doi: 10.1186/1471-2458-14-218.

[11] Xiangming Deng, Jinlong Teng, Xiucheng Nong, Bihan Yu, Liying Tang, Jinsong Liang, Zhuocheng Zou, Qiang Liu, Lu Zhou, Qirong Li, and Lihua Zhao. Characteristics of tcm constitution and related biomarkers for mild cognitive impairment. *Neuropsychiatric Disease and Treatment*, Volume 17: 1115–1124, 04 2021. doi: 10.2147/NDT.S290692.

[12] Mohammed Diykh, Yan Li, and Shahab Abdulla. Eeg sleep stages identification based on weighted undirected complex networks. *Computer Methods and Programs in Biomedicine*, 184:105116, 10 2019. doi: 10.1016/j.cmpb.2019.105116.

[13] Yuriko Doi, Masumi Minowa, Makoto Uchiyama, Masako Okawa, Keiko Kim, Kayo Shibui, and Yuichi Kamei. Psychometric assessment of subjective sleep quality using the japanese version of the pittsburgh sleep quality index

(psqi-j) in psychiatric disordered and control subjects. *Psychiatry research*, 97:165–72, 01 2001. doi: 10.1016/S0165-1781(00)00232-8.

[14] Shangyong Fan, Bin Chen, Xirui Zhang, Xiaojuan Hu, Lingshan Bao, Xiaodong Yang, Zhaobang Liu, and Yingcong Yu. Machine learning algorithms in classifying tcm tongue features in diabetes mellitus and symptoms of gastric disease. *European Journal of Integrative Medicine*, 43:101288, 2021. ISSN 1876-3820. doi: https://doi.org/10.1016/j.eujim.2021.101288. URL `https://www.sciencedirect.com/science/article/pii/S1876382021000068`.

[15] Toshi Furukawa, Norito Kawakami, Mari Oba, Yutaka Ono, Yoshibumi Nakane, Yosikazu Nakamura, Hisateru Tachimori, Noboru Iwata, Hidenori Uda, Hideyuki Nakane, Makoto Watanabe, Yoichi Naganuma, Yukihiro Hata, Masayo Kobayashi, Yuko Miyake, Tadashi Takeshima, and Takehiko Kikkawa. The performance of the japanese version of the k6 and k10 in the world mental health survey japan. *International journal of methods in psychiatric research*, 17:152–8, 09 2008. doi: 10.1002/mpr.257.

[16] Pei Gao, Zheng Chen, Ming Huang, Naoaki Ono, Shigehiko Kanaya, and MD Altaf-UI-Amin. An approach to construct and validate tcm dataset effective against bacterial pneumonia. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 102–103, 2021. doi: 10.1109/LifeTech52111.2021.9391949.

[17] Xuemin Gu, Guosheng Yin, and J. Lee. Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary clinical trials*, 36, 09 2013. doi: 10.1016/j.cct.2013.09.009.

[18] Yan Gu, Yifei Zhang, Xianzhe Shi, Xiaoying Li, Hong Jie, Jing Chen, Weiqiong Gu, Ning Wu, Guowang Xu, and Guang Ning. Effect of traditional chinese medicine berberine on type 2 diabetes based on comprehensive metabonomics. *Talanta*, 81:766–72, 05 2010. doi: 10.1016/j.talanta.2010.01.015.

[19] Hsiang-I Hou, Hsing-yu Chen, Jang-Jih Lu, Shih-Cheng Chang, Hsueh-Yu Li, Kun-Hao Jiang, and Jiun-Liang Chen. The relationships between leptin,

genotype, and chinese medicine body constitution for obesity. *Evidence-Based Complementary and Alternative Medicine*, 2021:1–11, 05 2021. doi: 10.1155/2021/5510552.

[20] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[21] Er-Yang Huan and Gui-Hua Wen. Multilevel and multiscale feature aggregation in deep networks for facial constitution classification. *Computational and Mathematical Methods in Medicine*, 2019:1–11, 12 2019. doi: 10.1155/2019/1258782.

[22] Er-Yang Huan and Gui-Hua Wen. Transfer learning with deep convolutional neural network for constitution classification with face image. *Multimedia Tools and Applications*, 79, 05 2020. doi: 10.1007/s11042-019-08376-5.

[23] Er-Yang Huan, Gui-Hua Wen, Shi-Jun Zhang, Dan-Yang Li, Yang Hu, Tian-Yuan Chang, Qing Wang, and Bing-Lin Huang. Deep convolutional neural networks for classifying body constitution based on face image. *Computational and Mathematical Methods in Medicine*, 2017:1–9, 10 2017. doi: 10.1155/2017/9846707.

[24] Shin-Yuan Hung, Yi-Cheng Ku, and Jui-Chi Chien. Understanding physicians' acceptance of the medline system for practicing evidence-based medicine: A decomposed tpb model. *International journal of medical informatics*, 81:130–42, 10 2011. doi: 10.1016/j.ijmedinf.2011.09.009.

[25] Akiomi Inoue, Norito Kawakami, Teruichi Shimomitsu, Akizumi Tsutsumi, Takashi Haratani, Toru Yoshikawa, Akihito Shimazu, and Yuko Odagiri. Development of a short version of the new brief job stress questionnaire. *Industrial health*, 52, 06 2014. doi: 10.2486/indhealth.2014-0114.

[26] Hajime Iwasa and Yuko Yoshida. Psychometric evaluation of the japanese version of ten item personality inventory (tipi-j) among middle-aged and elderly adults: Concurrent validity, internal consistency and test-retest reliability. *Cogent Psychology*, 5, 01 2018. doi: 10.1080/23311908.2018.1426256.

[27] Wang Ji. Research on constitution of chinese medicine and implementation of translational medicine. *Chin J Integr Med*, (2011):1–5, 2019. doi: 10.1007/s11655-014-2019-8.

[28] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1324–1330. International Joint Conferences on Artificial Intelligence Organization, 7 2020. URL `https://doi.org/10.24963/ijcai.2020/184`. Main track.

[29] Dihong JIANG, Ya-nan LU, Yu MA, and Yuanyuan WANG. Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement. *Expert Systems with Applications*, 121, 12 2018. doi: 10.1016/j.eswa.2018.12.023.

[30] QiaoYu Jiang, Jue Li, Guanghua Wang, and Jing Wang. The relationship between constitution of traditional chinese medicine in the first trimester and pregnancy symptoms: A longitudinal observational study. *Evidence-Based Complementary and Alternative Medicine*, 2016:1–8, 03 2016. doi: 10.1155/2016/3901485.

[31] Chen J Lai Y-Cheng J-Li F Xiao Y Jiang P-Sun X Luo R Zhao X-Liu Y Jing Y, Han S. Gut microbiota and urine metabonomics alterations in constitution after chinese medicine and lifestyle intervention. *The American Journal of Chinese Medicine*, 42:101288, 2021. ISSN 1-29. doi: 10.1142/S0192415X21500567.

[32] Ronald Kessler, P. Barker, Lisa Colpe, J. Epstein, J. Gfroerer, E. Hiripi, M. Howes, Sharon-Lise Normand, Ron Manderscheid, E. Walters, and Alan Zaslavsky. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32: 956–959, 01 2002.

[33] Daiki Koge, Naoaki Ono, Ming Huang, Altaf Amin, and Shigehiko Kanaya. Embedding of molecular structure using molecular hypergraph variational

autoencoder with metric learning. *Molecular informatics*, 40, 11 2020. doi: 10.1002/minf.202000203.

[34] Anjali Krishnan, Lynne Williams, Anthony McIntosh, and Hervé Abdi. Partial least squares (pls) methods for neuroimaging: A tutorial and review. *NeuroImage*, 56:455–75, 05 2011. doi: 10.1016/j.neuroimage.2010.07.034.

[35] Shih-hsien Kuo, Hsiao-ling Wang, Tzu-chi Lee, Te-fu Chan, Fan-hao Chou, Lih-mih Chen, and Wei-ting Lin. Traditional chinese medicine perspective on constitution transformations in perinatal women : A prospective longitudinal study. *Women and Birth*, pages 2–7, 2015. ISSN 1871-5192. URL `http://dx.doi.org/10.1016/j.wombi.2015.01.002`.

[36] Hyojeong Lee, Soo-Yong Shin, Myeongsook Seo, Gi-Byoung Nam, and Segyeong Joo. Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. *Scientific Reports*, 6:32390, 08 2016. doi: 10.1038/srep32390.

[37] Lingru Li, Haiqiang Yao, Ji Wang, Yingshuai Li, and Qi Wang. The role of chinese medicine in health maintenance and disease prevention: Application of constitution theory. *The American Journal of Chinese Medicine*, 47:1–12, 04 2019. doi: 10.1142/S0192415X19500253.

[38] Xiaojin Li, Licong Cui, Shiqiang Tao, Jing Chen, Xiang Zhang, and Guo-Qiang Zhang. Hyclasss: A hybrid classifier for automatic sleep stage scoring. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 02 2017. doi: 10.1109/JBHI.2017.2668993.

[39] Yung-Hui Li, Muhammad Saqlain Aslam, Kai-Lin Yang, Chung-An Kao, and Shin-You Teng. Classification of body constitution based on tcm philosophy and deep learning. *Symmetry*, 12(5), 2020. ISSN 2073-8994. doi: 10.3390/sym12050803.

[40] Chang Ke Li Beiting. Investigation and analysis on traditional chinese medicine constitution in 121 children with tic disorder in chengdu. *World Chinese Medicine*, 12(6):11–17, 2017.

[41] MA Wukai XIAO Lina-YANG Shufen LI Xingshu, LONG Yongjiao. Investigation on traditional chinese medicine constitution of patients with rheumatoid arthritis in guizhou province. *Chinese Journal of Integrative Nursing*, 6 (6):66, 2020.

[42] Yung-Cheng LIAO, Li-Li CHEN, Hsiao-Chiao WANG, Jui-Shan LIN, Tin-Kwang LIN, and Shu-Chuan LIN. The association between traditional chinese medicine body constitution deviation and essential hypertension: A case–control study. *Journal of Nursing Research*, Publish Ahead of Print, 06 2021. doi: 10.1097/JNR.0000000000000442.

[43] Jun-dai Lin, Li-li Chen, Jui-shan Lin, Chih-hung Chang, Yi-chia Huang, and Yi-chang Su. Bcq –: A body constitution questionnaire to assess part i : Establishment of a provisional version through a delphi process. *Forsch Komplementmed*, (91):234–241, 2012. doi: 10.1159/000343580.

[44] Jun-dai Lin, Jui-shan Lin, Li-li Chen, Chih-hung Chang, Yi-chia Huang, and Yi-chang Su. Bcqs : A body constitution questionnaire to assess stasis in traditional chinese medicine. *European Journal of Integrative Medicine*, pages 1–13, 2012. ISSN 1876-3820. doi: 10.1016/j.eujim.2012.05.001.

[45] WANG Yiping LIN Birong, LU Xingfeng. Investigation and analysis on distribution of traditional chinese medicine constitution among patients with primary hypertension in luzhou, china. *Chinese Journal of Integrative Nursing*, 5(8):45, 2019.

[46] Zhongyang Liu, Feifei Guo, Yong Wang, Chun Li, Xinlei Zhang, Honglei Li, Lihong Diao, Jiangyong Gu, Wei Wang, Dong Li, et al. Batman-tcm: a bioinformatics analysis tool for molecular mechanism of traditional chinese medicine. *Scientific reports*, 6(1):1–11, 2016.

[47] Hong JI Liyao ZHANG, Zhaoxiu WANG. A study on the current situation and correlation between the successful aging of the elderly in community and the constitution types of traditional chinese medicine. *Chinese Journal of Practical Nursing*, 34(32):2543–2547, 2018.

[48] QIU Chi LU Qiongfang. Clinical analysis of the pregnancy complications and outcomes in 125 women at advanced maternal age. *Chinese Journal of Integrative Nursing*, 5(2):93, 2019.

[49] LIU Wen-hong-ZHOU Dan-yang LIU Ying-hui HU Yong-bin MA Guo-ling SHOU Cheng-min CHEN Jia-wei MOU Xin, ZHAO Jin-xi. To discuss the susceptibility of constitution of tcm in diabetic nephropathy and the diversity of syndrome. *Journal of Basic Medicine in Traditional Chinese*, 2010.

[50] LIU Wen-hong-ZHOU Dan-yang LIU Ying-hui HU Yong-bin MA Guo-ling SHOU Cheng-min CHEN Jia-wei MOU Xin, ZHAO Jin-xi. Research on the effect of community diabetic prevented and treated by integrated traditional chinese and western medicine based on tcm constitution identification. *International Journal of Traditional Chinese Medicine*, 40(10):913–917, 2018.

[51] Susanne Mueller-Using, Torsten Feldt, Fred Sarfo, and Kirsten Eberhardt. Factors associated with performing tuberculosis screening of hiv-positive patients in ghana: Lasso-based predictor selection in a large public health data set. *BMC Public Health*, 16, 12 2016. doi: 10.1186/s12889-016-3239-y.

[52] Nobuyuki Nakatsu, Ryuichi Sawa, Shogo Misu, Yuya Ueda, and Rei Ono. Reliability and validity of the japanese version of the simplified nutritional appetite questionnaire in community-dwelling older adults. *Geriatrics and gerontology international*, 15, 12 2014. doi: 10.1111/ggi.12426.

[53] Atsushi Oshio, Shingo Abe, and Pino Cutrone. Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j). *The Japanese Journal of Personality*, 21:40–52, 01 2012. doi: 10.2132/personality.21.40.

[54] Atsuhiko Ota. Scientific base for the japanese stress check program. *Journal of Occupational Health*, 60, 11 2017. doi: 10.1539/joh.17-0288-ED.

[55] Mohamed I. Owis, Ahmed H. Abou-Zied, Abou Bakr M. Youssef, and Yasser M. Kadah. Study of features based on nonlinear dynamical modeling in ecg arrhythmia detection and classification. *IEEE Transactions on Biomedical Engineering*, 49:733–736, 2002.

[56] Birgit Prodinger, Alarcos Cieza, Cornelia Oberhauser, Jerome Bickenbach, Tevfik Ustun, Somnath Chatterji, and Prof. Dr. med. Gerold Stucki. Toward the international classification of functioning, disability and health (icf) rehabilitation set: A minimal generic set of domains for rehabilitation as a health strategy. *Archives of Physical Medicine and Rehabilitation*, 97, 01 2016. doi: 10.1016/j.apmr.2015.12.030.

[57] Wei Qu, Zhiyong Wang, Hong Hong, Zheru Chi, David Dagan Feng, Ron Grunstein, and Christopher Gordon. A residual based attention model for eeg based sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2833–2843, 2020. doi: 10.1109/JBHI.2020.2978004.

[58] Xiuxiu Sang, Zhongxia Wang, Shuyi Liu, and Ruilin Wang. Relationship between traditional chinese medicine(tcm) constitution and tcm syndrome in the diagnosis and treatment of chronic diseases. *Chinese Medical Sciences Journal*, 33(2):114–119, 2018. ISSN 1001-9294. doi: https://doi.org/10.24920/21806. URL `https://www.sciencedirect.com/science/article/pii/S1001929418300324`.

[59] Guang Shi, Hoko Kyo, Toshihiro Kawasaki, Shigehiko Kanaya, Mariko Sato, Saki Tokuda-Kakutani, Hiroshi Watanabe, Norihito Murayama, Minako Ohashi, Md Altaf-Ul-Amin, Naoaki Ono, Hiroki Tanaka, Satoshi Nakamura, Kazuo Uebaba, Nobutaka Suzuki, and Ming Huang. Evaluation and interpretation of 9 body constitution scores of ccmq-j by seven independent questionnaires. *Japanese Journal of Complementary and Alternative Medicine*, 16(2):79–93, 2019. doi: 10.1625/jcam.16.79.

[60] Huimei Shi, Yanbo Zhu, Qi Wang, Xiaohan Yu, Xiaomei Zhang, Lin Lin, and Li Shi. Factor analysis between health related quality of life and traditional chinese medicine constitution: lamination analysis on healthy adults and patients with chronic diseases in different subgroups. *Tianjin Journal of Traditional Chinese Medicine*, 35:251–254, 2018.

[61] YAO Shi-lin. From genetics to explore research thoughts on body constitution of chinese medicine. *Journal of Basic Medicine in Traditional Chinese*, 2010.

[62] WANG Bei SHI Tingting, TANG Shaohong. A study on the relationship between body mass index and traditional chinese medicine constitution classification. *Chinese Journal of Integrative Nursing*, 6(5):67, 2020.

[63] Youzhi Sun, Yi Zhao, Steve An Xue, and Jianping Chen. The theory development of traditional chinese medicine constitution: a review. *Journal of Traditional Chinese Medical Sciences*, 5(1):16–28, 2018. ISSN 2095-7548. doi: https://doi.org/10.1016/j.jtcms.2018.02.007. URL `https://www.sciencedirect.com/science/article/pii/S2095754817301679`.

[64] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP, 03 2017. doi: 10.1109/TNSRE.2017.2721116.

[65] Anna Lydia Svalastog, Doncho Donev, Nina Kristoffersen, and Srecko Gajovic. Concepts and definitions of health and health-related values in the knowledge landscapes of the digital society. *Croatian medical journal*, 58:431–435, 12 2017. doi: 10.3325/cmj.2017.58.431.

[66] David Tellez, Geert Litjens, Jeroen Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 08 2019. doi: 10.1109/TPAMI.2019.2936841.

[67] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 01 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[68] Fei Wang, Bo Wang, Long Wang, Zi-Yue Xiong, Wen Gao, Ping Li, and Hui-Jun Li. Discovery of discriminatory quality control markers for chinese herbal medicines and related processed products by combination of chromatographic analysis and chemometrics methods: Radix scutellariae as a case study. *Journal of Pharmaceutical and Biomedical Analysis*, 138, 02 2017. doi: 10.1016/j.jpba.2017.02.004.

[69] Ji Wang, Yingshuai Li, Cheng Ni, Huimin Zhang, Lingru Li, and Qi Wang. Cognition research and constitutional classification in chinese medicine. *The American journal of Chinese medicine*, 39:651–60, 01 2011. doi: 10.1142/S0192415X11009093.

[70] Q. Wang, Y.-B Zhu, H.-S Xue, and S. Li. Primary compiling of constitution in chinese medicine questionnaire. *Chinese Journal of Clinical Rehabilitation*, 10:12–14, 01 2006.

[71] Qi Wang. Classification and diagnosis basis of nine basic constitutions in chinese medicine. *J Beijing Univ Traditional Chinese Med*, 28:100029, 2005.

[72] Qi Wang. Individualized medicine , health medicine , and constitutional theory in chinese medicine. *Front. Med*, 6(1):1–7, 2012. doi: 10.1007/s11684-012-0173-y.

[73] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, PP:1–13, 09 2019. doi: 10.1109/TCYB.2019.2935141.

[74] Guihua Wen, Jiajiong Ma, Yang Hu, Huihui Li, and Lijun Jiang. Grouping attributes zero-shot learning for tongue constitution recognition. *Artificial Intelligence in Medicine*, 109:101951, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101951. URL `https://www.sciencedirect.com/science/article/pii/S0933365719310425`.

[75] Wendy Wong, Cindy Lo, Kuen Lam, Yi-chang Su, Sunny Jui-shan Lin, Eric Tat-chi Ziea, Vivian Taam, Lee Kin, Andrew Ka, and Lun Kwan. Measuring body constitution : Validation of the body constitution questionnaire (bcq) in hong kong. *Complementary Therapies in Medicine*, 22(4):670–682, 2014. ISSN 0965-2299. doi: 10.1016/j.ctim.2014.05.009.

[76] Han-kuei Wu, Yu-shien Ko, Yu-sheng Lin, Hau-tieng Wu, and Tung-hu Tsai. Complementary therapies in medicine the correlation between pulse diagnosis and constitution identification in traditional chinese medicine. *Comple-*

mentary Therapies in Medicine, 30(91):107–112, 2017. ISSN 0965-2299. doi: 10.1016/j.ctim.2016.12.005.

[77] S Yakubo, M Ito, Y Ueda, H Okamoto, Y Kimura, Y Amano, T Togo, H Adachi, T Mitsuma, and Kenji Watanabe. Pattern classification in kampo medicine. *Evidence-Based Complementary and Alternative Medicine*, 2014, 2014.

[78] Akihiro YAMAMORI, Hoko KYO, Tomoyuki WATANABE, Ming Huang, Naoaki Ono, Tetsuo Sato, Tetsuro ABE, Kazuo UEBABA, Katsushi KAWABATA, Keiho IMANISHI, Altaf Amin, Yanbo ZHU, Zhaoyu DAI, Qi WANG, Shigehiko Kanaya, Tomihisa Ohta, and Nobutaka SUZUKI. Relationship between 60 items in japanese version of the constitution in chinese medicine questionnaire (ccmq-j) based on multivariate analysis: Estimation of aging and bmi by ccmq-j scores. *Japanese Journal of Complementary and Alternative Medicine*, 13:43–56, 09 2016. doi: 10.1625/jcam.13.43.

[79] Hsien-Hui Yang, Chih-Sheng Chen, Hsin-Yi Lo, Chien-Yi Ho, Chia-Hung Kao, Tin-Yun Ho, and Tse-Yen Yang. The association between self-reported osteoporosis and chinese medicine-constitution questionnaire – a cross-sectional taiwan biobank study. 07 2020. doi: 10.21203/rs.3.rs-45021/v1.

[80] Zhaogeng Yang, Yanhui Li, Peijin Hu, Jun Ma, and Yi Song. Prevalence of anemia and its associated factors among chinese 9-, 12-, and 14-year-old children: Results from 2014 chinese national survey on students constitution and health. *International Journal of Environmental Research and Public Health*, 17(5), 2020. ISSN 1660-4601. doi: 10.3390/ijerph17051474. URL `https://www.mdpi.com/1660-4601/17/5/1474`.

[81] Takahashi K Yoshimura Y. Excel eiyho-kun (nutrition) food frequency questionnaire based on food group ffqg (computer manual and software). *Tokyo*, 4:3, 2014.

[82] Ye Yuan and Wei Liao. Design and implementation of the traditional chinese medicine constitution system based on the diagnosis of tongue and consultation. *IEEE Access*, 9:4266–4278, 2021. doi: 10.1109/ACCESS.2020.3047452.

[83] Junming Zhang, Ruxian Yao, Wengeng Ge, and Jinfeng Gao. Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel eeg. *Computer Methods and Programs in Biomedicine*, 183: 105089, 09 2019. doi: 10.1016/j.cmpb.2019.105089.

[84] Yuan Zhang, Yao Guo, Po Yang, Wei Chen, and Benny Lo. Epilepsy seizure prediction on eeg using common spatial pattern and convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*, 24(2):465–474, 2020. doi: 10.1109/JBHI.2019.2933046.

[85] ZHOU Chen ZHUANG Ai-wen YANG Dong-mei ZHANG Yi, ZHANG Wei. The relationship about the depression in late-pregnancy and constitution in traditional chinese medicine. *Journal of Basic Medicine in Traditional Chinese*, 2010.

[86] Xinyu Zhao, Xinggui Tan, Hongfei Shi, and Daozong Xia. Nutrition and traditional chinese medicine (tcm): a system's theoretical perspective. *European Journal of Clinical Nutrition*, 75:1–7, 09 2020. doi: 10.1038/s41430-020-00737-w.

[87] Xiaolin Zhou, Hongxia Ding, Wanqing Wu, Yuanting Zhang, and Alena Talkachova. A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate. In *PloS one*, 2015.

[88] Y.-B Zhu, Q. Wang, and H. Origasa. Evaluation on reliability and validity of the constitution in chinese medicine questionnaire (ccmq). *Zhongguo Xing Wei Yi Xue Ke Xue*, 16:651–654, 01 2007.

[89] Yanbo ZHU, Hideki Origasa, Kazuo UEBABA, Fenghao XU, and Qi WANG. Development and validation of the japanese version of the constitution in chinese medicine questionnaire (ccmq). *Kampo Medicine*, 59:783–792, 01 2008. doi: 10.3937/kampomed.59.783.

[90] YB ZHU. Development of the constitution in chinese medicine questionnaire (ccmq). *Japanese Journal of Public Health*, 52:383. URL `https://ci.nii.ac.jp/naid/10024040616/en/`.