Social Medicine and Health Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/socmhm.2024.050219
ISSN 2616-2210 Vol. 5 Num. 2

Correlation Analysis of Hypertension Risk with Diet and Lifestyle Habits: Based on Random Forest and SHAP Model

Yuhang Zhong[1,a,#], Haiyuan Nong[2,b,#], Yuxin Liu[2,c], Guorui Zhao[2,d,*]

[1]School of Materials Science and Engineering, Guangdong Ocean University, Yangjiang, Guangdong, China
[2]School of Computer Science and Engineering, Guangdong Ocean University, Yangjiang, Guangdong, China
[a]m2471477868@163.com, [b]oceanedgen@gmail.com, [c]13427937113@163.com, [d]zhaoguoruimath@foxmail.com
[*]Corresponding author
[#]Equal contributions

Keywords: Hypertension risk, Eating habits, Living habits, Random forest, SHAP model

Abstract: Hypertension has become a significant health threat for residents. Investigating the relationship between disease risk and dietary and lifestyle habits is both theoretically important and practically valuable. This paper examines the connection between the risk of hypertension and the dietary and lifestyle choices of residents, using cross-sectional data from an epidemiological survey conducted in Shenzhen. To start, we developed an index system that integrates the Chinese Dietary Guidelines for residents with guidelines for managing hypertension. Next, we preprocessed the data using the 3σ criterion and a sample equalization method. We then constructed a random forest model to explore the relationship between the risk of hypertension and diet and lifestyle factors. To improve model interpretability, we applied the newly developed SHAP model for quantitative analysis. The results indicated that six factors were most strongly associated with the risk of hypertension: work intensity, milk intake, high-protein food consumption, exercise intensity, the weekly incidence of skipping breakfast, and vegetable intake. Among these, the interaction between high-protein food intake and work intensity was particularly significant.

1. Introduction

Health is a fundamental and universal need for all human beings, symbolizing national prosperity. Modern lifestyle shifts have led to chronic NCDs, notably in China, including hypertension, cardiovascular/cerebrovascular diseases, diabetes, malignancies, and COPD. Hypertension, prevalent and risky, often leads to cardiovascular/cerebrovascular diseases and severe complications like stroke, myocardial infarction, heart failure, and chronic kidney disease, causing high disability and mortality rates. This burdens medical resources and the economy [1]. Thus, in-

depth studies and effective hypertension prevention/control measures are crucial. Optimizing diagnostics and treatment plans is vital for mitigating risks, alleviating medical burdens, and enhancing patients' quality of life.

The Chinese Guidelines for the Prevention and Treatment of Hypertension (2024 Revised Edition) define hypertension as SBP ≥140 mmHg or DBP ≥90 mmHg without medication. Initially seen as physiological, hypertension's risks have become clearer with medical research. Many studies aim to identify hypertension risk factors but rely on traditional linear models, which struggle with complex relationships and have limited analysis capabilities. Given hypertension's multiple influences and intricate nonlinear relationships, these models fail to fully explain risk factor connections.

Machine learning models overcome traditional limitations by exploring nonlinear relationships in hypertension risks. However, interpretability issues limited their use. Advances in explanatory methods, especially the SHAP model, now demystify machine learning. SHAP calculates influence magnitude, direction, and variable importance, and analyzes interactions, enhancing its applicability.

The main work of this paper is reflected in two aspects. First, based on indicators such as the Dietary Guidelines for Chinese Residents (2022) Balanced Diet Guidelines [2] and the hypertension treatment guidelines of the WHO/ISH Hypertension Conference, a multi-dimensional and hierarchical grading index system of diet and living habits was constructed. Secondly, the random forest model and SHAP model were used to model and analyze the diet and lifestyle data, and the machine learning model was used to better handle the nonlinear relationship and complex data structure to reveal the correlation between the risk of hypertension and lifestyle and dietary habits. Based on the results, suggestions to reduce the risk of hypertension were provided.

2. Literature review

Research on hypertension risk factors has garnered global attention. In China, Xie et al. (2013) studied how diet and exercise affect the quality of life of elderly hypertensive patients through community nursing [3]. Sun (2001) found misunderstandings about dietary structure among hypertensive patients and highlighted the need for health education [4]. Internationally, the European Society of Hypertension (ESH) and others have issued guidelines summarizing evidence to aid professionals in managing hypertension [5]. The 1998 WHO/ISH Hypertension Conference in Japan finalized hypertension treatment guidelines, emphasizing non-drug therapies like diet, exercise, psychology, and environmental stress, highlighting the impact of lifestyle on hypertension risk [6].

Traditional statistical methods often miss nonlinear relationships and complex data in hypertension studies. Recently, machine learning has offered new possibilities. Zhang Lina et al. (2006) used Logistic regression to screen factors [7]. Shi Ye-Wen(2024) further screened predictors with univariate Logistic regression and built a hypertension risk model using multivariate Logistic regression with Lasso [8]. Jingjie Wu and Lili Yang (2021) compared the prediction performance of different machine learning and statistical methods, verifying predictors like age, gender, and obesity [9].

In international research, Wu et al. (2020) selected 9 outcome events of young hypertensive patients to build XGBoost, COX regression, and Framingham risk score models. Studies showed machine learning outperformed traditional methods [10]. Adeleke O et al. (2024) used machine learning algorithms to predict hypertension, confirming its potential [11]. Yong Whi Jeong et al. (2022) used a machine learning model for South Korean Public Health data (2002-2017), finding XGBoost superior to logistic regression [12]. Chang et al. (2019) built a hypertension prediction model based on multiple machine-learning comparisons, considering data leakage [13].

A large number of studies have shown that hypertension is closely related to lifestyle and dietary habits, and can be prevented and controlled by corresponding measures. However, the interaction and influencing mechanism between these factors have not been discussed in detail. The existing studies also mainly focus on the influence of a single factor and ignore the interaction between multiple factors, and most of them use traditional statistical methods, which cannot fully consider nonlinear relationships and complex data structures. Therefore, this paper will use a random forest model to analyze the relationship between the risk of hypertension and diet and lifestyle, and then use the SHAP model to rank the relationship between the factors and analyze the interaction between the Features.

3. Index System Construction and Data Description

3.1. Data Sources

This study utilized data from Shenzhen Health Research Department's questionnaire on "Epidemiology of Chronic Non-Communicable Diseases and Related Factors." It collected 7709 questionnaires, covering demographics (height, weight, age), dietary habits (food intake, frequency), work and exercise intensities, and physiological indicators (diastolic and systolic blood pressure). With a relatively balanced male-to-female ratio (3840 females, 2783 males), questionnaires with logical contradictions or numerical abnormalities were excluded based on the 3σ criterion. Ultimately, 6623 valid questionnaires were selected for analysis.

3.2. Hypertension Measurement Criteria and Development of Index System

To study the relationship between the risk of hypertension and diet and lifestyle habits, it is necessary to understand how to measure hypertension. According to the Chinese Guidelines for the Prevention and Treatment of Hypertension (2024 Revised Edition), hypertension refers to a chronic disease with systolic blood pressure ≥140mmHg and/or diastolic blood pressure ≥90mmHg without antihypertensive drugs. Based on this standard, this paper uses the systolic blood pressure and diastolic blood pressure data of the participants in the questionnaire to judge whether they have hypertension.

Based on the "Balanced Diet Guidelines for Chinese Residents (2022)", eleven indicators of dietary habits were extracted from the questionnaire, including the intake (g) of vegetables, fruits, milk, whole grain, Root crops, high-protein food, oil, salt, and alcohol, daily type of food intake, and weekly incidence of missing breakfast.

To comprehensively evaluate the impact of lifestyle on hypertension, five lifestyle indicators were established based on the WHO/ISH guidelines on hypertension treatment and other literature and data: weekly cigarette count, cumulative exercise time (min/week), work intensity, leisure or household intensity, and exercise intensity. The basic information index of sex was retained in the construction of the index system.

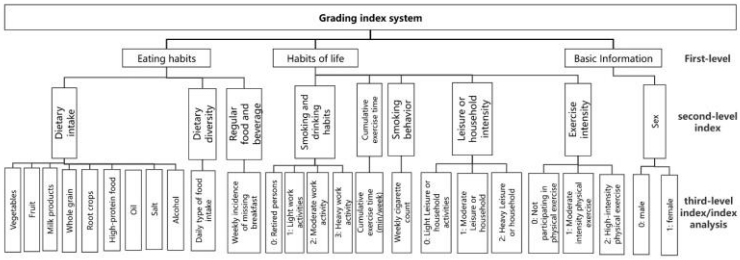

Figure 1: Grading index system.

A total of 17 indicators provided a comprehensive data basis for subsequent analysis. The specific index system and normal values are shown in Figure 1.

3.3. Data Description and Processing

The mean value and coefficient of variation were used to describe the indicators of dietary habits, including the intake of vegetables, fruits, milk products, whole grain, root crops, high-protein food, oil, salt, and alcohol, daily type of food intake, weekly incidence of missing breakfast, the Cumulative exercise time (min/week), and Weekly cigarette count. The higher the coefficient of variation, the more scattered the data distribution of this feature. Conversely, a smaller coefficient of variation indicates a more concentrated data distribution. For the other indicators, this paper counts their values and uses this information to understand the differences in the living habits of different groups. The number of people with and without hypertension was also counted here. The mean and coefficient of variation are shown in Table 1.

Table 1: Mean and coefficient of variation.

Features	Mean value	Coefficient of variation
Vegetable	280.46	0.52
Fruit	135.11	0.78
Milk products	80.40	1.32
Whole grain	343.55	0.41
Root crops	8.08	1.19
High-protein food	218.30	0.53
Oil	34.56	0.49
Salt	4.18	0.57
Alcohol	50.03	3.11
Daily type of food intake	7.35	0.29
Weekly incidence of missing breakfast	0.69	2.54
Cumulative exercise time (min/week)	75.77	1.54
Weekly cigarette count	15.63	2.52
Features	The distribution	
Hypertension	No hypertension: 5711	
	having hypertension: 512	
Sex	male：2783	
	female：3840	
Work intensity	Retired persons: 917	
	Light work activities: 4045	
	Moderate work activity: 1222	
	Heavy work activity: 39	
Leisure or household intensity	Light Leisure or household activities: 2720	
	Moderate Leisure or household: 3471	
	Heavy Leisure or household: 32	
Exercise intensity	Not participating in physical exercise:2977	
	Moderate intensity physical exercise: 2364	
	High-intensity physical exercise: 882	

Table 1's coefficient of variation shows large individual differences in alcohol (3.11) and cigarette consumption (2.52), indicating varied smoking and drinking habits. Food category/day had a low coefficient (0.29), suggesting similar daily food variety. Most participants had light-moderate work and leisure, with few reporting heavy intensities, preferring a relaxed lifestyle. Some didn't exercise, possibly due to lower work and leisure intensities.

At the same time, the number of hypertension patients was 512, accounting for 8.23% of the total, the proportion was low, and the data was not evenly distributed, which was not conducive to the follow-up study. Therefore, the sample was balanced, and the data of 12172 were finally obtained for analysis by over-sampling method.

4. Introduction of Research Methods and Model Construction

4.1. SHAP Model

SHAP model (Shapely Additive Explanations) [14] is a post-hoc explanatory model based on the additivity of Shapely value [15] in cooperative game theory. Its core idea lies in calculating the marginal contribution of features to the output of the model, to explain the results of the model.

The SHAP model views all features as contributors and calculates their marginal contributions across all feature sequences, averaging to obtain the SHAP value. This value indicates the feature's contribution degree and direction to the model's prediction.

In machine learning, features interact, affecting predictions jointly. This interaction is crucial for understanding models and improving predictions. The SHAP model, based on Shapely values, considers joint contributions of multiple features and calculates their average SHAP interaction attribution value.

The SHAP interaction attribution value resembles the SHAP value, reflecting contribution degree (magnitude) and direction (positive/negative) under multi-feature interactions.

Traditional machine learning explanatory models can only understand feature contributions, not specific feature value impacts. The SHAP model resolves this. Common SHAP value algorithms are TreeSHAP [16] and KernelSHAP. KernelSHAP is slow but versatile, applicable to all models. This paper uses a tree model (RF) and selects TreeSHAP for faster computation.

4.2. Random Forest Model Construction and Parameter Setting

The random forest algorithm uses bootstrap sampling to create N training sets (2/3 of the original data) and builds N CART decision trees. Each tree selects M features ($m \leq M$) and splits nodes based on the minimum Gini coefficient. The final prediction is determined by voting among the N trees. The remaining 1/3 of the data (Out-of-bag) is used for internal error estimation, yielding the OOB error [17].

This paper used 17 features (diet, living habits, gender) as independent variables and hypertension as the dependent variable. The training set was 70%, with 100 decision trees, Gini splitting, and a max tree depth of 10. Model performance was evaluated using accuracy, precision, recall, and F1 score.

5. Analysis of Empirical Results

5.1. Evaluation of Model Effect

This paper compared the performance of four machine learning models (RF, LR, SVM, BP) using 17 features as input. RF demonstrated superior performance with an AUC of 0.998 (95% CI: 0.997 to 0.999; $P<0.001$), significantly higher than the other models: LR (AUC: 0.685; 95%CI: 0.668-0.702; $P<0.001$), SVM (AUC: 0.522; 95%CI: 0.503-0.540; $P<0.001$), and BP (AUC: 0.690; 95%CI: 0.673-0.707; $P<0.001$). The ROC curve is depicted in Figure 2.

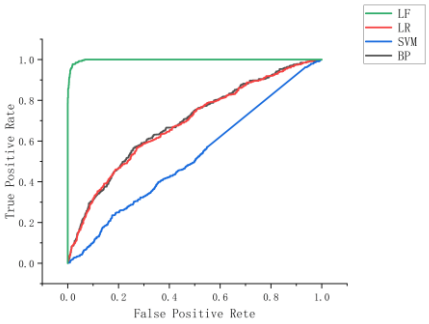

Figure 2: ROC curve.

Additionally, RF also outperformed the other models in terms of accuracy, recall, precision, and F1-Score. See Table 2 for details.

Therefore, the selection of RF as the model for this study was justified based on its superior performance across multiple evaluation metrics, including the AUC value.

Table 2: The test set model evaluated the results.

	RF	LR	SVM	BP
Rate of accuracy	96.40%	63.10%	50.50%	64.40%
Rate of recall	100%	63.10%	50.50%	64.40%
Rate of precision	93.10%	63.20%	50.50%	64.50%
F1-Score	0.964	0.631	0.504	0.644

5.2. Main Empirical Conclusions Based on the SHAP Model

After comparing multiple machine learning models, the Random Forest (RF) model was selected, and the SHAP model was used to analyze it, exploring factor-hypertension risk relationships and interactions.

SHAP values were calculated and ranked, revealing the top six factors affecting hypertension risk: work intensity (0.098), milk intake (0.058), high-protein food intake (0.039), exercise intensity (0.033), weekly incidence of missing breakfast (0.031), and vegetable intake (0.026). (see Figure 3).

Secondly, the PDP diagram (as shown in Figure 4) was used to explore the influence of six main factors on disease risk. Key findings include:

Work intensity: Retirees (intensity 0) have a lower hypertension risk, while those with light, moderate, and heavy work activities (intensities 1, 2, and 3) face a higher risk.

Milk intake: Increased milk consumption is associated with a higher hypertension risk.

High-protein food intake: Below 200g, hypertension risk rises rapidly with intake; above 200g, the risk continues to increase but at a slower rate.

Physical exercise: Moderate exercise reduces hypertension risk, while no exercise or vigorous exercise increases it.

Skipping breakfast: Skipping breakfast zero times per week significantly lowers hypertension risk, whereas skipping 1-7 times per week increases it.

Vegetable intake: Below 400g, increased vegetable consumption effectively reduces hypertension risk; above 400g, however, the risk rises with intake.

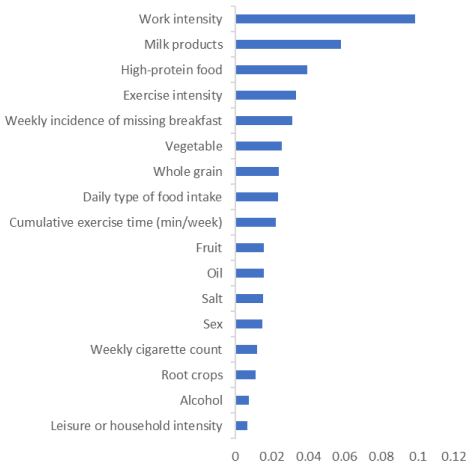

Figure 3: Mean (|SHAP value|).

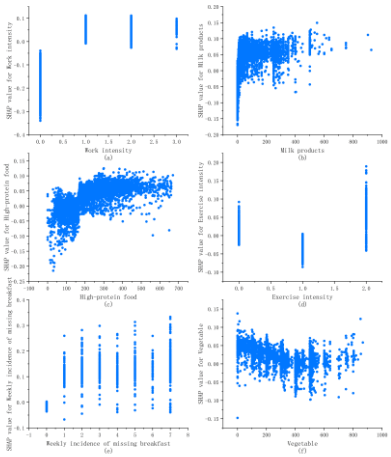

Figure 4: Dependence plot.

Finally, considering the multifaceted nature of hypertension and the interconnectedness of its contributing factors, this study delves into the interactions among these features. Utilizing the SHAP model's interaction attribution value, a heatmap matrix (as shown in Figure 5) is constructed to visualize these interactions. Notably, the analysis highlights a significant interaction between work intensity and high-protein food intake, with an attribution value of 0.0089, indicating the strongest link among the features studied. Consequently, this research focuses specifically on exploring the interplay between work intensity and high-protein food intake.

The dependence scatter plot of work intensity and high-protein food intake was plotted (as shown in Figure 6). Figure 6 shows that increased high-protein food intake raises hypertension risk for both retirees and workers of various intensities. The risk increase was slower for retirees but faster for those with lighter, moderate, and high-intensity work.

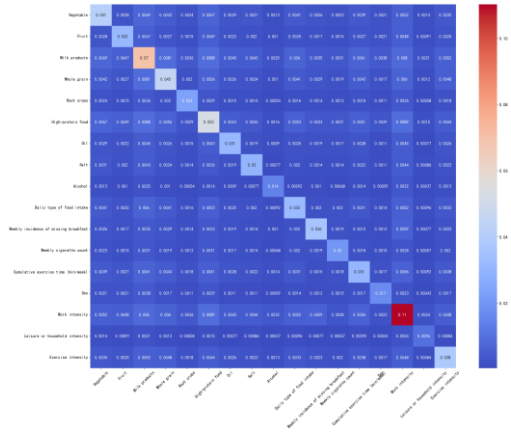

Figure 5: SHAP interaction values matrix.

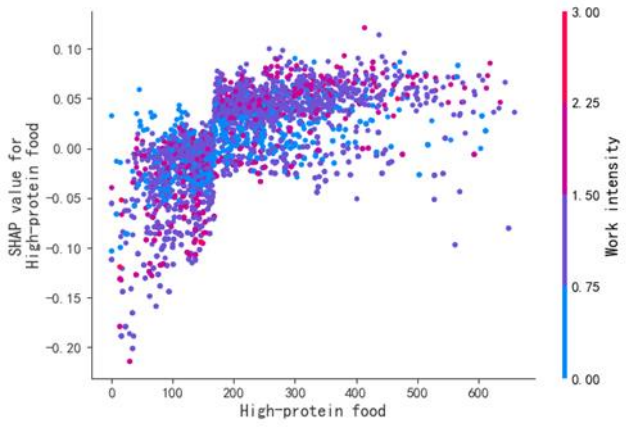

Figure 6: Dependence scatter plot.

6. Conclusions and Recommendations

To investigate hypertension risk factors among Shenzhen residents, this paper established a multi-dimensional index system based on an epidemiological survey. Machine learning models were applied, with the random forest model chosen for its predictive accuracy. Further analysis was conducted using the SHAP model to explore the relationship between hypertension risk and diet, and living habits.

The results indicated that diet and living habits significantly impact hypertension risk. Key

factors included work intensity, milk and high-protein food intake, exercise intensity, missing breakfast, and vegetable intake. Higher work intensity, milk and protein-rich food consumption, and missing breakfast increased risk. Moderate exercise and vegetable intake reduced risk but effectiveness diminished beyond a certain point. Notably, a synergistic effect between high-protein food intake and work intensity was observed, exacerbating hypertension risk.

Based on empirical findings, reducing work intensity, milk and high-protein food intake, ensuring regular breakfasts, controlling vegetable intake, and engaging in moderate exercise can effectively lower hypertension risk.

This paper was primarily authored by Yuhang Zhong and Haiyan Nong, with Zhong responsible for data collection and analysis, and Nong focusing on literature review and manuscript composition.

References

[1] The Chinese Hypertension Prevention and Control Guideline Revision Committee, the Hypertension League (China), the Chinese Medical Healthcare International Exchange Promotion Association Hypertension Branch, et al. Chinese Hypertension Prevention and Control Guideline (2024 Revised Edition) [J]. Chinese Journal of Hypertension (Chinese and English), 2024, 32(07): 603-700.

[2] Chinese Dietary Guidelines for Residents (2022): Balanced Diet Criteria [J]. Disease Prevention and Control Bulletin, 2024, 39(01):95.

[3] Xie Jianxiu, Chen Shuiqiong, Ye Cuihua, Hu Yanping. Analysis of the impact of community nursing intervention on the quality of life of elderly patients with hypertension [J]. Contemporary Medicine, 2013, 19(09):119-120.

[4] Sun Yanhong. A dietary survey and dietary guidance for 176 patients with hypertension [J]. Journal of Nursing, 2001(08):451-453.

[5] Mancia, G., & Grassi, G. (Eds.). Manual of Hypertension of the European Society of Hypertension (2nd ed.). CRC Press. 2014.

[6] Lin Jinxiu, Wu Kegui. Guidelines for the treatment of hypertension by the World Health Organization/International Society of Hypertension in 1999 [J]. Journal of Hypertension, 1999(02):4-7.

[7] Zhang Lina, Chen Jian'er, Zhang Tao, et al. Investigation on the prevalence characteristics and related factors of hypertension [J]. Chinese Journal of Public Health, 2006(01):93-94.

[8] Shi Ye-Wen, Ma Li-Na, Zhu Si-Min, et al. Construction of OSA-related hypertension prediction model based on line chart [J]. Journal of Clinical Otolaryngology - Head and Neck Surgery, 2024, 38(11): 1024-1030+1037.

[9] Wu Jingjie, Yang Lili. Research progress of machine learning in building hypertension risk model [J]. Nursing and Rehabilitation, 2019, 20(02):33-36.

[10] Wu X, Yuan X, Wang W, et al. Value of a Machine Learning Approach for Predicting Clinical Outcomes in Young Patients With Hypertension. Hypertension. 2020; 75(5):1271-1278.

[11] Adeleke O ,Adebayo S ,Aworinde H , et al. Machine learning evaluation of a hypertension screening program in a university workforce over five years.[J]. Scientific reports, 2024, 14(1):30255.

[12] Jeong, Yong Whi; Jung, Yeojin. Prediction Model for Hypertension and Diabetes Mellitus Using Korean Public Health Examination Data (2002–2017) [J]. Diagnostics (Basel), 2022, Aug.

[13] Chang, W. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data [J]. Diagnostics, 2019, 9: 178. DOI: 10.3390/diagnostics9040178.

[14] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30. 2017.

[15] Shapley, Lloyd S., A Value for N-Person Games. Santa Monica, CA: RAND Corporation, 1952.

[16] S. Lundberg, G. Erion, S. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv, 2018.

[17] Zhang Xiaoyu, Li Fengri, Zhen Zhen, et al. Forest vegetation classification of land satellite-8 remote sensing images based on random forest model [J]. Journal of Northeast Forestry University, 2016, 44(06): 53-57+74.