Explainable analysis of the Relationship between Hypertension with Gas leakages

Khongorzul Dashdondov[1], Kyuri Jo[2], Mi-Hye Kim[2]*
[1]Researcher, Department of Computer Engineering, Chungbuk National University, Chungbuk 28644, Korea
[2]Professor, Department of Computer Engineering, Chungbuk National University, Chungbuk 28644, Korea

설명 가능한 인공지능 기술을 활용한 가스누출과 고혈압의 연관 분석

홍고르출 [1], 조겨리 [2], 김미혜 [2]*
[1] 충북대학교 컴퓨터공학과 박사, [2] 충북대학교 컴퓨터공학과 교수
khongorzul63@gmail.com, kyurijo@chungbuk.ac.kr, mhkim@cbnu.ac.kr

요 약

Hypertension is a severe health problem and increases the risk of other health issues, such as heart disease, heart attack, and stroke. In this research, we propose a machine learning-based prediction method for the risk of chronic hypertension. The proposed method consists of four main modules. In the first module, the linear interpolation method fills missing values of the integration of gas and meteorological datasets. In the second module, the OrdinalEncoder-based normalization is followed by the Decision tree algorithm to select important features. The prediction analysis module builds three models based on k-Nearest Neighbors, Decision Tree, and Random Forest to predict hypertension levels. Finally, the features used in the prediction model are explained by the DeepSHAP approach. The proposed method is evaluated by integrating the Korean meteorological agency dataset, natural gas leakage dataset, and Korean National Health and Nutrition Examination Survey dataset. The experimental results showed important global features for the hypertension of the entire population and local components for particular patients. Based on the local explanation results for a randomly selected 65-year-old male, the effect of hypertension increased from 0.694 to 1.249 when age increased by 0.37 and gas loss increased by 0.17. Therefore, it is concluded that gas loss is the cause of high blood pressure.

1. INTRODUCTION

In this article, we propose a machine learning-based risk prediction method for hypertension using feature selection and corresponding analysis by integrating the Korean meteorological agency dataset, natural gas leakage dataset, and Korean National Health and Nutrition Examination Survey (KNHANES) dataset from the Korea Centers for Disease Control and Prevention, as shown in Figure 1. The paper has been divided into four modules. First, we fill in missing data based on the linear interpolation method on the integrated dataset of gas and meteorological datasets. Then select essential features using the Random Forest (RF) classifier with OrdinalEncoder (OE)-based normalization and correlations with corresponding analysis. The third module uses three algorithms, k-Nearest Neighbors (kNN), Decision Tree (DT) and RF to predict hypertension level. The final module explains the features used in the predictive models based on DeepSHAP method.

2. METHODOLOGY

This study used the linear interpolation method and substitution of the mean value to replace missing values in the environmental data set. Then we normalize data by the OE technique. Behind the normalization, we select essential features using the RF classifier.

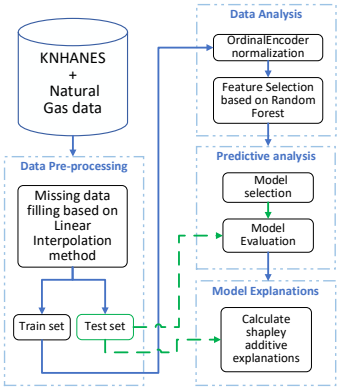
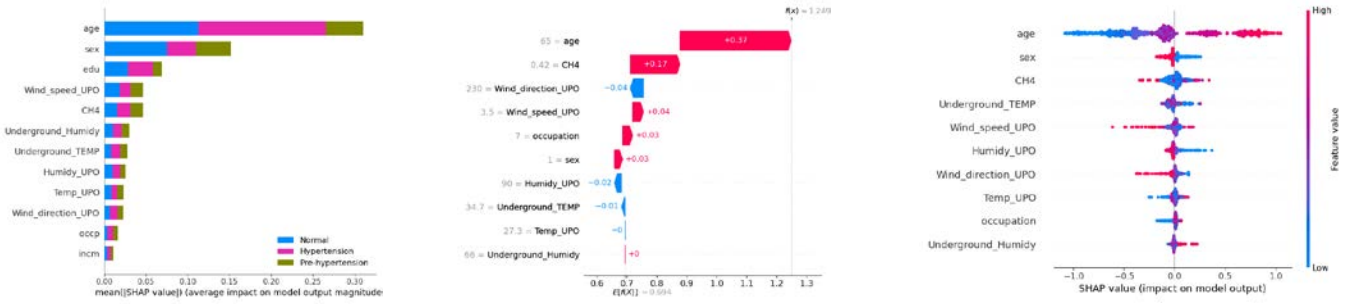


Figure1. System architecture for proposed model.



(a) (b) (c)

Figure 2. The global and local explanation for hypertension among Koreans from the environment and population-based approach

The gas data was collected by UPO company on a trial basis from August 1 to August 31, 2020, in Jeollanamdo province, Korea. Initially, we integrated UPO dataset with Korean meteorological dataset and then filled a row of missing values using the linear interpolation method. Then we added location environment weather data from the Korean meteorological web application. After that, we combined it with the KNHANES dataset. It consists of a health examination of various diseases, health interviews, and nutrition surveys of the Korean population. We have analyzed samples duration the years 2016-2018. We generated the target value for the upper 19-year-old hypertension patients. After that, we also determined the correlation with hypertension based on correspondence analysis. We make train machine learning models for predictive analysis in this labeled dataset. After the train, we tested predictive and evaluating models by accuracy measurement.

Then indicates that the proposed features selection-based method is suitable for predicting the risk of hypertension detection for combined datasets. Studies show that hypertension is related to air pollution and social, demographic, and health factors.

3. RESULTS AND DISCUSSIONS

The features were arranged in descending order of overall importance ratings in the proposed model development to provide a model explanation. The prediction model was developed with 12 features as a result of the XGboost classifier's decision. Figure 2(a) shows the most essential 12 global features regarding the relationship between hypertension in the general population and gas leakage in Korea, along with their significance ratings. For the prediction model of risk of hypertension across the entire population, age, sex, edu (education level), Wind_speed_UPO, CH4 (natural gas), Underground_TEMP, Underground_Humidy, Temp_UPO, Humidy_UPO, Wind_direction_UPO, occupation, income were maintained most significant features with importance scores of 0.401, 0.079, 0.079, 0.071, 0.070, 0.066, 0.055, 0.053, 0.048, 0.039, 0.024, and 0.015, respectively.

High-scored features are at the top of the list because of the DeepSHAP-based global explanation technique. These significance scores are appropriate for comprehending the total population sample, but not at the individual level. Additionally, while some factors appear to have less influence on the hypertension prediction model throughout the entire sample, they may have a key impact on some patients for identifying hypertension.

In Figure 2(b) and 2(c), the local explanation of the randomly chosen individual is shown. The local explanation result exhibits the most important 10 features of hypertension and environment, negative risk factors were colored by blue, and positive risk factors were marked by red. A negative coefficient indicates that the occurrence becomes less likely as the predictor increases. The randomly selected individual had the highest favorable connections for reducing hypertension: age = 65, sex = 1. (male), and CH4 (natural gas) = 0.42 (medium). On the contrary, Humidy_UPO = 90, Wind_direction_UPO = 230, and Underground_Temp = 34.7 were significantly negatively associated with preventing hypertension. This randomly selected man had age of 0.37 and increased gas leakage by 0.17 the effect of hypertension increased from 0.694 to 1.249. Therefore, it is concluded that gas loss is the cause of hypertension.

Acknowledgment

This research was financially supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2022-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

[1] Khongorzul, D., Kim M.-H., Lee S. M. (2019). OrdinalEncoder based DNN for Natural Gas Leak Prediction. J. Korea Convergence Society, 10(10), 7-13.

[2] Dashdondov, K. and Kim, M.H., 2021. Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction. Neural Processing Letters, pp.1-13.

[3] Available website: UPO company, http://www.upokorea.com/new/pdf/UPO_Catalogue.pdf

[4] Available website: Korean public data portal https://www.data.go.kr/dataset/15000099/openapi.do