Medical & healthcare, Smart device, AI

Development of a machine learning model for high-performance prediction of lifestyle-related disease risk using medical big data

Health and Counseling Center

Specially Appointed Assistant Professor Asuka Oyama
Researchmap https://researchmap.jp/a.ooyama?lang=en

Visiting Researcher Hiroe Seto
Researchmap https://researchmap.jp/SetoHiroe?lang=en

Professor Ryohei Yamamoto
Researchmap https://researchmap.jp/yamamotor?lang=en

Abstract

Lifestyle-related diseases can be prevented by improving one's lifestyle. However, improving and maintaining a healthy lifestyle is not easy. Recognizing the risk of lifestyle-related diseases and enhancing health awareness is crucial to improve one's lifestyle. Various disease prediction models, mainly machine learning models, have been developed to estimate the risk of lifestyle-related disease onset in individuals. However, the amount of data available during training was small, and ensuring sufficient accuracy and reliability was difficult. We aimed to develop an AI model that could predict the probability of diabetes mellitus (DM), hypertension, and dyslipidemia onset within three years with high performance, using the big data from the specific health checkups of the Osaka Prefectural National Health Insurance.

Background & Results

It is difficult for residents without specialized medical knowledge to understand the items in the specific health checkups. This study showed the superiority of machine learning for big data analysis. Furthermore, we have successfully developed an AI model that can predict the probability of lifestyle-related disease onset with high accuracy (fig. 1). The model has now been implemented in the mobile healthcare app "Asmile." Users can access the probabilities of lifestyle-related disease onset through this app (fig. 2). It is expected to be an essential tool to improve health awareness and promote autonomous health behavior.

Significance of the research and Future perspective

Previous studies have suggested no difference in the predictive accuracy between classical statistical and AI models in disease onset prediction. This is because it is difficult to obtain and utilize large-scale real-world data, and sufficient data are lacking to maximize the capabilities of machine learning models. With the cooperation of the Osaka Prefectural Health Insurance Federation, we used big data from specific health checkups of 600,000 individuals annually. From this dataset, we analyzed a subset of approximately 280,000 individuals without a history of DM. We evaluated the number of individuals used for model training and the accuracy of the probability estimation for developing DM. We used a logistic regression model as the statistical model and a gradient boosting decision tree model (LightGBM) as the machine learning model, known for its flexibility and learning speed. As shown in Figure 1, although there was almost no difference in prediction accuracy between the two models for a dataset of approximately 1,000 individuals, the prediction accuracy of LightGBM improved as the number of individuals in the training dataset increased. Using a dataset of 10,000 individuals, LightGBM outperformed logistic regression in terms of prediction accuracy. This analysis suggests that machine learning models outperform statistical models when using big data.

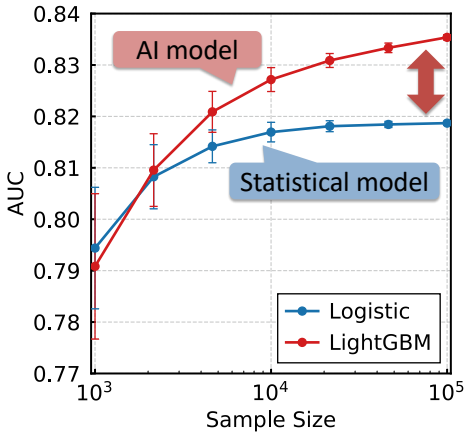

Fig. 1. Relationship between the number of data used for training and the AUC.

Fig. 2. The lifestyle-related disease onset prediction model implemented in the mobile healthcare app "Asmile".

Patent

Treatise

Seto, Hiroe; Oyama, Asuka; Kitora, Shuji et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Scientific Reports. 2022, 12, 15889. doi: 10.1038/s41598-022-20149-z
Seto, Hiroe; Toki, Hiroshi; Kitora, Shuji et al. Seasonal variations of the prevalence of metabolic syndrome and its markers using big-data of health checkups. Environmental Health and Preventive Medicine. 2024, 29, 2. doi: 10.1265/ehpm.23-00216
Kotoku, Jun'ichi; Oyama, Asuka; Kitazumi, Kanako et al. Causal relations of health indices inferred statistically using the DirectLiNGAM algorithm from big data of Osaka prefecture health checkups. PLoS ONE. 2020, 15 (12), e0243229. doi: 10.1371/journal.pone.0243229
Oyama, Asuka; Kotoku, Jun'ichi; Toki, Hiroshi et al. High-mortality society seen by medical big data: Deciphering the super aging society from Osaka KDB big data. Information Processing Society of Japan. 2024, 65(4), e15–e23. doi: 10.20729/00233422 (In Japanese)

URL

Keyword specific health checkup, real-world data, machine learning, lifestyle-related disease, national health insurance