Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240512

Machine Learning-Based Predictive Models for Early Detection of Cardiovascular Diseases: A Study Utilizing Patient Samples from a Tertiary Health Promotion Center in Korea

Kanghyuck LEE[a,b], Seol Whan OH[a,b], Sung-Hwan KIM[c] , Taehoon KO[a,b,1] and In Young CHOI[a,b,2]
[a] *Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea*
[b] *Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea*
[c] *Department of Internal Medicine, Division of Cardiology, Seoul St. Mary's Hospital College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea*
ORCiD ID: Kanghyck Lee https://orcid.org/0009-0006-4257-4776
Seol Whan Oh https://orcid.org/0000-0002-0328-9634
Sung-Hwan Kim https://orcid.org/0000-0001-6805-0416
Taehoon Ko https://orcid.org/0000-0002-4045-0036
In Young Choi https://orcid.org/0000-0002-2860-9411

Abstract. A machine learning model was developed for cardiovascular diseases prediction based on 21,118 patient checkups data from a tertiary medical institution in Seoul, Korea, collected between 2009 and 2021. XGBoost algorithm showed the highest predictive performance, with an average AUROC of 0.877. In survival analysis, XGBSE achieved an AUROC exceeding 0.9 for 2-9 year predictions, with a C-index of 0.878 across all diseases, outperforming Cox regression (C-index of 0.887). A high-performance prediction model for cardiovascular diseases using the XGBSE algorithm was successfully developed and is poised for real-world clinical application following external simplification and validation.

Keywords. Artificial Intelligence, Electronic Health Records, Survival Analysis

1. Introduction

In South Korea, the healthcare system performs annual health checkups for the entire population. These health checkup data are very useful because they include health screenings and questionnaires. This study aimed to develop Machine learning (ML)-

[1] Corresponding Author: Taehoon Ko; E-mail: thko@catholic.ac.kr.
[2] Corresponding Author: In Young Choi; E-mail: iychoi@catholic.ac.kr.

based predictive models for hypertension, dyslipidemia, and ischemic heart disease using health checkup data from a medical institution in Seoul, Korea.

2. Methods

The study utilized examination and questionnaire data from 21,118 patient samples collected between 2009 and 2021 at the Health Promotion Center of a tertiary medical institution in Seoul, Korea (Table 1). For predicting the occurrence of CVDs, logistic regression, decision tree, random forest, and eXtreme Gradient Boosting (XGBoost) [1,2] were used to compare the performance metric area under the receiver operating characteristic curve (AUROC). Additionally, to predict the incidence rate over time, survival analysis was performed using Cox regression and XGBoost Survival Embedding (XGBSE), with the Concordance index (C-index) evaluated for each model.

Table 1. The Number of utilized variables

Total 63 variables				
Health screening		Health questionnaires		
Physical examination	Laboratory findings	Demographics	Past history	Lifestyle habits
12 variables	10 variables	6 variables	32 variables	3 variables

3. Results and Discussion

The XGBoost model had the highest average AUROC of 0.877. In survival analysis, XGBSE showed very few differences from Cox regression (C-index of 0.878), with an average AUROC above 0.9 for all diseases and a C-index of 0.887 for the 2-9 year prediction. XGBSE demonstrated superior performance, particularly in survival analysis, suggesting its potential for accurate CVDs prediction over extended periods. Adjusting the threshold improved sensitivity without sacrificing predictive accuracy, highlighting the utility of ML techniques in developing robust CVDs prediction models.

4. Conclusions

This study researched to develop a model for predicting CVDs using health checkup data [3]. It was found that the XGBSE outperformed other existing models. With external validation and simplification, it has a high potential for real-world clinical application and can facilitate tailored CVDs prevention interventions for at-risk individuals.

References

[1] Shin J, Lee J, Ko T, Lee K, Choi Y, Kim HS. Improving machine learning diabetes prediction models for the utmost clinical effectiveness. J Pers Med. 2022;12(11):1899.
[2] Barnwal A, Cho H, Hocking T. Survival regression with accelerated failure time model in XGBoost. J Comput Graph Stat. 2022;31(4):1292-1302.
[3] Kim YT, Lee WC, Cho B. National screening program for the transitional ages in Korea. Journal of the Korean Medical Association. 2010;53(5):371-376.