


Nonclinical Features in Predictive Modeling of Cardiovascular Diseases: A Machine Learning Approach

Mirza Rizwan Sajid¹ · Noryanti Muhammad¹  · Roslinazairimah Zakaria¹ · Ahmad Shahbaz² · Syed Ahmad Chan Bukhari³ · Seifedine Kadry⁴ · A. Suresh⁵

¹ Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang Darul Makmur, Malaysia

² Punjab Institute of Cardiology, Lahore 54000, Pakistan

³ Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. Johns University, New York, NY 11439, USA

⁴ Faculty of Applied Computing and Technology, Noroff University College, Kristiansand, Norway

⁵ Department of Computer Science and Engineering, SRM Institute of Science & Technology, Kattankulathur, Chengalpattu (D.t) 603 203, Tamilnadu, India

ABSTRACT

Background In the broader healthcare domain, the prediction bears more value than an explanation considering the cost of delays in its services. There are various risk prediction models for cardiovascular diseases (CVDs) in the literature for early risk assessment. However, the substantial increase in CVDs-related mortality is challenging global health systems, especially in developing countries. This situation allows researchers to improve CVDs prediction models using new features and risk computing methods. This study aims to assess nonclinical features that can be easily available in any healthcare systems, in predicting CVDs using advanced and flexible machine learning (ML) algorithms.

Methods A gender-matched case–control study was conducted in the largest public sector cardiac hospital of Pakistan, and the data of 460 subjects were collected. The dataset comprised of eight nonclinical features. Four supervised ML algorithms were used to train and test the models to predict the CVDs status by considering traditional logistic regression (LR) as the baseline model. The models were validated through the train–test split (70:30) and tenfold cross-validation approaches.

Results Random forest (RF), a nonlinear ML algorithm, performed better than other ML algorithms and LR. The area under the curve (AUC) of RF was 0.851 and 0.853 in the train–test split and tenfold cross-validation approach, respectively. The nonclinical features yielded an admissible accuracy (minimum 71%) through the LR and ML models, exhibiting its predictive capability in risk estimation.

Conclusion The satisfactory performance of nonclinical features reveals that these features and flexible computational methodologies can reinforce the existing risk prediction models for better healthcare services.

Keywords Nonclinical features, Cardiovascular diseases, Machine learning algorithms, Risk prediction models, Cost-effective model

