ClinicSearch

International Journal of Cardiovascular Medicine

Amit Mandal *

Open Access

Research Article

Unraveling Coronary Artery Disease Risk Factors: Insights from Machine Learning and Statistical Analysis

Alexander A. Huang [1,2]*, Samuel Y. Huang [1]

[1]Cornell University

[2]Northwestern University Feinberg School of Medicine

***Corresponding Author:** Alexander A. Huang, Cornell University, Northwestern University Feinberg School of Medicine.

Received date: February 15, 2024; Accepted date: February 29, 2024; Published date: March 12, 2024

Citation: Alexander A. Huang, Samuel Y. Huang, (2024), Unraveling Coronary Artery Disease Risk Factors: Insights from Machine Learning and Statistical Analysis, *International Journal of Cardiovascular Medicine*, 3(2); **DOI:**10.31579/2834-796X/058

Abstract

Coronary artery disease (CAD) stands as a significant health challenge worldwide, necessitating comprehensive efforts in risk factor identification to improve prevention and management strategies.(1) Huang and Huang present a pioneering study utilizing machine learning techniques to explore risk factors associated with CAD

Keywords: coronary artery disease; risk factors; ecg

1. Introduction

Coronary artery disease (CAD) stands as a significant health challenge worldwide, necessitating comprehensive efforts in risk factor identification to improve prevention and management strategies.(1) Huang and Huang present a pioneering study utilizing machine learning techniques to explore risk factors associated with CAD.(1) By harness
ing the extensive National Health and Nutrition Examination Survey (NHANES) dataset, the authors aimed to unravel intricate risk factor interactions and assess their predictive capabilities using transparent machine learning methodologies. (1)

2. Methods

Employing a retrospective, cross-sectional cohort design, the study delved into NHANES data spanning 2017 to 2020. (1) Participants who completed demographic, dietary, exercise, and mental health questionnaires and provided laboratory and physical exam data were included. (1) Initially, univariate logistic models were employed to discern significant covariates linked to CAD. (1) Subsequently, the XGBoost machine learning algorithm, renowned for its accuracy in healthcare prediction, was applied.(1) Covariates were then ranked based on their contribution to the model's prediction, and Shapely Additive Explanations (SHAP) were utilized for visualization and **interpretation of risk factor relationships.(1)**

3. Results

The study encompassed 7,929 participants, among whom 4.5% were diagnosed with CAD. Impressively, the XGBoost model exhibited robust predictive accuracy (AUROC = 0.89). Notable predictors identified included age, total cholesterol, total platelets, and family history of heart attack. The SHAP visualizations corroborated these findings, revealing nuanced relationships between these factors and CAD risk, aligning closely with existing literature. Furthermore, non-linear associations were observed for cholesterol and platelet count, highlighting the need for nuanced risk factor analysis.

4. Discussion

The study's findings underscore the potential of machine learning in unraveling the complex interplay of demographic, physiological, and lifestyle factors in predicting CAD risk.(1) Transparent methodologies such as SHAP facilitate the interpretation and validation of model predictions, enhancing confidence in the identified risk factors.(1) Despite the retrospective design and reliance on self-reported data, the study benefits from the inclusion of a large, demographically diverse NHANES cohort, which bolsters the generalizability and replicability of the findings.(1)

5. Conclusion

Huang and Huang's study represents a significant advancement in CAD risk prediction, shedding light on key risk factors and their relative contributions.(1) The identified predictors, including age, cholesterol, platelet count, and family history, underscore the multifaceted nature of CAD etiology.(1) The study's transparent approach and use of SHAP visualizations provide valuable insights for clinical practice and public health interventions, paving the way for personalized medicine and targeted interventions in CAD management.(1)

6. Future Directions and Implications for Practice

Various statistical methods were employed in Huang and Huang's study to elucidate risk factors for coronary artery disease (CAD).(2-4) Univariate logistic regression was initially utilized to identify covariates significantly associated with CAD, based on their p-values.(5-8) This method allowed for the exploration of individual risk factors and their respective contributions to CAD prediction. Subsequently, the XGBoost machine learning algorithm was employed, known for its robust performance in healthcare prediction

tasks; XGBoost operates by iteratively improving the predictive accuracy of an ensemble of decision trees, effectively capturing complex interactions among covariates.(5-8) Furthermore, SHAP (Shapely Additive Explanations) was utilized to visualize the relationships between covariates and CAD risk, providing insights into the direction and magnitude of their effects.(7, 9-12) These statistical methods collectively enabled a comprehensive assessment of risk factors for CAD, from individual associations to nuanced interactions, contributing to a deeper understanding of disease etiology.(13-16)

Moving forward, continued research is warranted to advance our understanding of CAD risk factors and their implications for clinical practice.(1, 17, 18) Longitudinal studies tracking the progression of identified risk factors and their impact on CAD outcomes could provide valuable insights into disease trajectories and inform targeted interventions.(19-24) Additionally, collaboration between data scientists, clinicians, and public health professionals is crucial for translating machine learning insights into actionable strategies for CAD prevention and management.(18, 20, 21, 23, 25-27)

The study's findings have significant implications for clinical practice, emphasizing the importance of personalized risk assessment in CAD management. (1, 16, 18, 19, 28, 29) Clinicians can leverage machine learning-based risk prediction tools to stratify patients based on individualized risk profiles, enabling tailored interventions and optimizing patient outcomes.(1, 16, 18, 19, 28, 29) By identifying high-risk individuals earlier and implementing targeted preventive measures, healthcare providers can mitigate the burden of CAD and improve population health outcomes.(17, 21, 23, 24, 27, 30)

As machine learning continues to evolve, the methodologies employed in Huang and Huang's study offer valuable insights for future research directions. One avenue for advancement lies in the refinement and optimization of machine learning algorithms for healthcare prediction tasks. While XGBoost demonstrated robust performance in CAD risk prediction, exploring alternative algorithms and ensemble techniques could further enhance predictive accuracy and generalizability.(8, 31-35) Future research may also focus on developing interpretable and transparent machine learning models, akin to SHAP visualizations, to facilitate model validation and ensure clinical relevance.(13, 15, 34-38)

Moreover, incorporating diverse and comprehensive datasets, akin to NHANES, holds promise for enriching machine learning models and uncovering novel insights into disease etiology. Integrating multi-modal data sources, including genomics, imaging, and wearable sensor data, could provide a more holistic understanding of disease mechanisms and enable personalized risk assessment.(39-42) Additionally, leveraging advanced data preprocessing techniques, such as feature engineering and dimensionality reduction, can help alleviate data sparsity and improve model performance, particularly in scenarios with high-dimensional data.(43-46)

Another crucial aspect for future research is the integration of machine learning models into clinical decision support systems (CDSS) and healthcare workflows.(2, 4, 5, 8, 10) Collaborative efforts between data scientists, clinicians, and healthcare stakeholders are essential for developing user-friendly and clinically actionable tools.(3, 47) Emphasizing interpretability and transparency in model outputs can foster trust and acceptance among healthcare professionals, facilitating the adoption of machine learning-driven approaches in real-world settings.(3, 12, 16, 47, 48) Moreover, ongoing evaluation and validation of CDSS in clinical practice are imperative to ensure safety, efficacy, and adherence to regulatory standards.(7, 11, 49, 50)

Furthermore, addressing ethical and regulatory considerations is paramount in the deployment of machine learning models in healthcare. Future research must prioritize ethical guidelines, privacy protection, and data security to safeguard patient rights and ensure responsible use of sensitive healthcare data.(7, 11, 49-52) Moreover, fostering interdisciplinary collaborations and establishing robust governance frameworks can promote transparency, accountability, and equity in machine learning-driven healthcare initiatives.(53-56)

In conclusion, Huang and Huang's study exemplifies the potential of machine learning methodologies in advancing healthcare research and clinical practice. By embracing interdisciplinary collaboration, leveraging diverse datasets, and prioritizing transparency and ethical considerations, future research endeavors can further propel the integration of machine learning into healthcare systems.(4, 6, 8, 57) These efforts hold promise for revolutionizing disease prevention, diagnosis, and treatment, ultimately improving patient outcomes and advancing population health.

7. Limitations

While the study demonstrates notable strengths, including its transparent methodology and utilization of a large dataset, several limitations warrant consideration. The retrospective design and reliance on self-reported data introduce inherent biases and potential inaccuracies. (1, 23, 26-28) Furthermore, the NHANES cohort's voluntary nature may lead to selection bias, limiting the generalizability of the findings.(16, 17, 22, 26, 58) Future studies employing prospective designs and automated data collection methods could mitigate these limitations and provide more accurate assessments of CAD risk factors.(21-25, 28)

8. Conclusion

In conclusion, Huang and Huang's study represents a seminal contribution to CAD research, leveraging machine learning techniques to uncover key risk factors and their predictive capabilities. Despite its limitations, the study provides valuable insights into the multifactorial nature of CAD etiology and underscores the potential of machine learning in enhancing risk prediction and informing clinical practice. Continued research in this field holds promise for advancing personalized medicine and mitigating the global burden of coronary artery disease.

References

1. Huang AA, Huang SY. (2023). Use of machine learning to identify risk factors for coronary artery disease. PLoS One. 18(4):e0284103.
2. Rashijane LT, Mokoena K, Tyasi TL. (2023). Using Multivariate Adaptive Regression Splines to Estimate the Body Weight of Savanna Goats. Animals (Basel). 13(7).
3. Materka A, Jurek J. (2024). Using Deep Learning and B-Splines to Model Blood Vessel Lumen from 3D Images. Sensors (Basel). 24(3).
4. Bucher A, Genest C, Lockhart RA, Neslehova JG. (2023). Asymptotic behavior of an intrinsic rank-based estimator of the Pickands dependence function constructed from B-splines. Extremes (Boston). 26(1):101-38.
5. Mandeville JB, Efthimiou N, Weigand-Whittier J, Hardy E, Knudsen GM, Jorgensen LM, et al. (2024). Partial volume correction of PET image data using geometric transfer matrices based on uniform B-splines. Phys Med Biol. 69(5).
6. Gauthier J, Wu QV, Gooley TA. (2023). Correction: Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transplant. 58(8):962.
7. Bekar Adiguzel M, Cengiz MA. (2023). Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria. Heliyon. 9(9):e19964.

8. Alavi J, Aminikhah H. (2023). Orthogonal cubic splines for the numerical solution of nonlinear parabolic partial differential equations. MethodsX. 10:102190.
9. Sadrara M, Khorrami MK. (2023). Principal component analysis-multivariate adaptive regression splines (PCA-MARS) and back propagation-artificial neural network (BP-ANN) methods for predicting the efficiency of oxidative desulfurization systems using ATR-FTIR spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc. 300:122944.
10. Mushtaq K, Zou R, Waris A, Yang K, Wang J, Iqbal J, et al. (2023). Multivariate wind power curve modeling using multivariate adaptive regression splines and regression trees. PLoS One. 18(8):e0290316.
11. Dyrting S, Taylor A. (2023). Estimating age-specific mortality using calibrated splines. Popul Stud (Camb). 1-18.
12. Bach NH, Vu LH, Nguyen VD, Pham DP. (2023). Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational auto-encoder. Sci Rep. 13(1):19984.
13. Yang Y, Chi Y, Yuan S, Zhang Q, Su L, Long Y, et al. (2022). The relationship between ventilatory ratio (VR) and 28-day hospital mortality by restricted cubic splines (RCS) in 14,328 mechanically ventilated ICU patients. BMC Pulm Med. 22(1):229.
14. Wawrzyniak J. (2022). Methodology for Quantifying Volatile Compounds in a Liquid Mixture Using an Algorithm Combining B-Splines and Artificial Neural Networks to Process Responses of a Thermally Modulated Metal-Oxide Semiconductor Gas Sensor. Sensors (Basel). 22(22).
15. Sun F, Cai Z. (2024). A Family of Generalized Cardinal Polishing Splines. IEEE Trans Image Process. 33:1952-64.
16. Huang AA, Huang SY. (2023). Quantification of the Effect of Vitamin E Intake on Depressive Symptoms in United States Adults Using Restricted Cubic Splines. Curr Dev Nutr. 7(2):100038.
17. Huang AA, Huang SY. (2023). Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. PLoS One. 18(2):e0281922.
18. Huang AA, Huang SY. (2023). Use of machine learning to identify risk factors for insomnia. PLoS One. 18(4):e0282622.
19. Huang AA, Huang SY. (2023). Computation of the distribution of model accuracy statistics in machine learning: Comparison between analytically derived distributions and simulation-based methods. Health Sci Rep. 6(4):e1214.
20. Huang AA, Huang SY. (2023). Dendrogram of transparent feature importance machine learning statistics to classify associations for heart failure: A reanalysis of a retrospective cohort study of the Medical Information Mart for Intensive Care III (MIMIC-III) database. PLoS One. 18(7):e0288819.
21. Huang AA, Huang SY. (2023). Technical Report: Machine-Learning Pipeline for Medical Research and Quality-Improvement Initiatives. Cureus. 15(10):e46549.
22. Huang AA, Huang SY. (2023). Use of feature importance statistics to accurately predict asthma attacks using machine learning: A cross-sectional cohort study of the US population. PLoS One. 18(11):e0288903.
23. Huang AA, Huang SY. (2023). Covariate dependent Markov chains constructed with gradient boost modeling can effectively generate long-term predictions of obesity trends. BMC Res Notes. 16(1):346.
24. Huang AA, Huang SY. (2023). Stochastic modeling of obesity status in United States adults using Markov Chains: A nationally representative analysis of population health data from 2017-2020. Obes Sci Pract. 9(6):653-60.
25. Huang AA, Huang SY. (2023). Hospitalized COVID-19 patients with diabetes have an increased risk for pneumonia, intensive care unit requirement, intubation, and death: A cross-sectional cohort study in Mexico in 2020. Health Sci Rep. 6(4):e1222.
26. Huang AA, Huang SY. (2023). Quantification of the Relationship of Pyridoxine and Spirometry Measurements in the United States Population. Curr Dev Nutr. 7(8):100078.
27. Huang AA, Huang SY. (2023). Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension. J Clin Hypertens (Greenwich). 25(12):1135-44.
28. Huang AA, Huang SY. (2023). Diabetes is associated with increased risk of death in COVID-19 hospitalizations in Mexico 2020: A retrospective cohort study. Health Sci Rep. 6(7):e1416.
29. Huang AA, Huang SY. (2023). Exploring Depression and Nutritional Covariates Amongst US Adults using Shapely Additive Explanations. Health Sci Rep. 6(10):e1635.
30. Huang AA, Huang SY. (2023). Increased vigorous exercise and decreased sedentary activities are associated with decreased depressive symptoms in United States adults: Analysis of The National Health and Nutrition Examination Survey (NHANES) 2017-2020. Health Sci Rep. 6(8):e1473.
31. Xu Y, Han D, Xu F, Shen S, Zheng X, Wang H, et al. (2022). Using Restricted Cubic Splines to Study the Duration of Antibiotic Use in the Prognosis of Ventilator-Associated Pneumonia. Front Pharmacol. 13:898630.
32. Wakamiya A, Nagase S, Kusano K. (2024). Successful release of multiple splines of a multipolar catheter entrapped in a mechanical mitral valve: video presentation of a validated method. Eur Heart J Case Rep. 8(3): ytae091.
33. Tian Y, Baro E, Zhang R. (2019). Performance evaluation of regression splines for propensity score adjustment in post-market safety analysis with multiple treatments. J Biopharm Stat. 29(5):810-21.
34. Sahs J, Pyle R, Damaraju A, Caro JO, Tavaslioglu O, Lu A, et al. (2022). Shallow Univariate ReLU Networks as Splines: Initialization, Loss Surface, Hessian, and Gradient Flow Dynamics. Front Artif Intell. 5:889981.
35. Wang J, Bi S, Liu W, Zhou L, Li T, Macleod I, et al. (2023). Stitching Locally Fitted T-Splines for Fast Fitting of Large-Scale Freeform Point Clouds. Sensors (Basel). 23(24).
36. Tirink C, Eyduran E, Faraz A, Waheed A, Tauqir NA, Nabeel MS, et al. (2021). Use of multivariate adaptive regression splines for prediction of body weight from body measurements in Marecha (Camelus dromedaries) camels in Pakistan. Trop Anim Health Prod. 53(3):339.
37. Saadaoui F, Khalfi M. (2022). Revisiting Islamic banking efficiency using multivariate adaptive regression splines. Ann Oper Res.1-29.
38. Rodriguez-Alvarez MX, Durban M, Eilers PHC, Lee DJ, Gonzalez F. (2023). Multidimensional adaptive P-splines with application to neurons' activity studies. Biometrics. 79(3):1972-85.
39. Scope Crafts E, Lu H, Ye H, Wald LL, Zhao B. (2022). An efficient approach to optimal experimental design for magnetic resonance fingerprinting with B-splines. Magn Reson Med. 88(1):239-53.
40. Qin X, Hung J, Knuiman MW, Briffa TG, Teng TK, Sanfilippo FM. (2020). Comparison of medication adherence measures derived from linked administrative data and associations with

mortality using restricted cubic splines in heart failure patients. Pharmacoepidemiol Drug Saf. 29(2):208-18.
41. Peters J. (2020). Refinable tri-variate C (1) splines for box-complexes including irregular points and irregular edges. Comput Aided Geom Des. 80.
42. Onak O, Erenler T, Serinagaoglu Y. (2022). A Novel Data-Adaptive Regression Framework Based on Multivariate Adaptive Regression Splines for Electrocardiographic Imaging. IEEE Trans Biomed Eng. 69(2):963-74.
43. Riesenfeld RF, Johnson C, Kasik D, M CW. (2022). The Development of B-Splines for CAD. IEEE Comput Graph Appl. 42(2):90-100.
44. Peters-Sanders L, Sanders H, Goldstein H, Ramachandran K. (2023). Using Multivariate Adaptive Regression Splines to Predict Lexical Characteristics' Influence on Word Learning in First Through Third Graders. J Speech Lang Hear Res. 66(2):589-604.
45. Munoz-Osorio GA, Tirink C, Tyasi TL, Ramirez-Bautista MA, Cruz-Tamayo AA, Dzib-Cauich DA, et al. (2024). Using fat thickness and longissimus thoracis traits real-time ultrasound measurements in Black Belly ewe lambs to predict carcass tissue composition through multiresponse multivariate adaptive regression splines algorithm. Meat Sci. 207:109369.
46. Meng C, Yu J, Chen Y, Zhong W, Ma P. (2022). Smoothing splines approximation using Hilbert curve basis selection. J Comput Graph Stat. 31(3):802-12.
47. Lamichhane BP. (2023). A mixed finite element discretisation of linear and nonlinear multivariate splines using the Laplacian penalty based on biorthogonal systems. MethodsX. 10:101962.
48. Chevremont W. (2023). SpatDistCalib: a GUI Python software for spatial-distortion correction of 2D detectors using splines. J Appl Crystallogr. 56(Pt 3):860-7.
49. Kaya H, Hardy DJ, Skeel RD. (2021). Multilevel summation for periodic electrostatics using B-splines. J Chem Phys. 154(14):144105.
50. Marquez M, Meza C, Lee DJ, De la Cruz R. (2023). Classification of longitudinal profiles using semi-parametric nonlinear mixed models with P-Splines and the SAEM algorithm. Stat Med. 42(27):4952-71.
51. Gogel B, Welham S, Cullis B. (2022). Empirical comparison of time series models and tensor product penalised splines for modelling spatial dependence in plant breeding field trials. Front Plant Sci. 13:1021143.
52. Donohue MC, Langford O, Insel PS, van Dyck CH, Petersen RC, Craft S, et al. (2023). Natural cubic splines for the analysis of Alzheimer's clinical trials. Pharm Stat. 22(3):508-19.
53. Gascoigne C, Smith T. (2023). Penalized smoothing splines resolve the curvature identifiability problem in age-period-cohort models with unequal intervals. Stat Med. 42(12):1888-908.
54. Elhakeem A, Hughes RA, Tilling K, Cousminer DL, Jackowski SA, Cole TJ, et al. (2022). Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. BMC Med Res Methodol. 22(1):68.
55. Davidson SE, Wheeler MW, Auerbach SS, Sivaganesan S, Medvedovic M. (2022). ALOHA: Aggregated local extrema splines for high-throughput dose-response analysis. Comput Toxicol. 21.
56. D'Urso P, De Giovanni L, Vitale V. (2022). Spatial robust fuzzy clustering of COVID 19 time series based on B-splines. Spat Stat. 49:100518.
57. Carta-Bergaz A, Avila P, Arenal A. (2023). PentaRay floppy splines for coronary artery ostia location. Rev Esp Cardiol (Engl Ed).
58. Huang A, Henderson G, Profeta A, Pfeiffer M, Feinstein LH, deLahunta M, et al. (2023). Lack of compensation of energy intake explains the success of alternate day feeding to produce weight loss. Physiol Behav. 263:114128.