Abstract:

Cardiac arrest and cardiovascular diseases (CVD) are one of the main causes of mortality in the world, which is aggravated by smoking and other risk factors. The current literature review brings together several papers that used machine learning (ML) as their predictor in CVD risk using various data types: Heart rate variability (HRV) in smokers, electronic health records (EHRs) in the long-term, and hospital-based clinical records. These models, like 1D-CNN, Random Forest (RF), and deep learning, consistently performed better than the traditional ones, scoring up to 94.98 percent and an AUC of up to 0.865. The major ones are overfitting, heterogeneity, limited generalizability, and the absence of clinical integration. These results demonstrate that ML can be used to stratify personalized and timely risks early in advance, and require the development of standardized validation and interpretation to be applied in the real world

Literature Review of existing work:

Cardiovascular diseases (CVD) and cardiac arrest is one of the most important public health burdens and it is found that smoking is one of the most significant modifiable risk factors that increase the speed of clot formation, increase heart rate and decrease oxygen supply, which contributes to 5-6 lakh deaths every year in India and 7 million worldwide (Raja et al., 2024). Raja et al. (2024) used a publicly available dataset of 1,562 middle-aged subjects (811 smokers were at risk of cardiac arrest, 751 non-smokers) and 19 HRV features in the hemodynamic, time, frequency, and non-linear domains in a focused study. In the min-max scaled and

80:20 train and test split evaluation, four ML models were studied: Support Vector Machine (SVM), Artificial Neural Network (ANN), Deep Neural Network (DNN), and one-dimensional Convolutional Neural Network (1D-CNN). 1D- CNN was the best performer, with an accuracy of 94.98, sensitivity of 97.39, and specificity of 92.73, which outperformed DNN (94.25%), ANN (93.29%), and SVM (89.45%), which overfitted because of higher-dimensional physiological data (Raja et al., 2024). In line with this, Ali and Ahmed (2025) performed a retrospective study of 300 patient datasets of Shar Hospital, Iraq, and included 10 clinical risk factors (e.g., age, smoking, diabetes, family history). Three models, Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) were tested in different ratios of train to test and RF showed a higher average score of 90.78% in the 10 splits (and 88.94% in the 20 trial) and a strong performance with relation to classification through confusion matrices, whereas importance of features made age and family history viable predictors (Ali and Ahmed, 2025).

Liu et al. (2024) have conducted a PRISMA-conformant systematic review and meta-analysis of 20 studies (2010-2024) with 32 ML models and 26 statistical models (conventional) based on EHR data that were synthesized with 5-10-year CVD risk prediction. Random Forest produced the largest pooled AUC of 0.865 (95% CI: 0.8120917), then deep learning, 0.847 (95% CI: 0.766927), and, significantly, the conventional scores such as the QRISK3 and ASCVD, 0.765 (95% CI: 0.7347927) (Liu et al., 2024). Even though there was severe discrimination, high heterogeneity (I 2 over 99 percent), bias during publication, inconsistent hyperparameter coverage, and scanty external validation, clinical applicability was limited (Liu et al., 2024). The three studies all found ensemble and deep learning-based methods, especially RF and convolutional architectures to be superior in dealing with complex, heterogeneous biomedical data without the intensive feature engineering required by traditional methods although 1D-CNN outperforms RF in sequential HRV signal detection and RF offers more interpretability by being able to rank features in importance (Raja et al., 2024; Ali and Ahmed, 2025; Liu et al., 2024).

The overall evidence advocates ML-driven change in the prediction of CVD risk, allowing real-time and personalized interventions that might decrease mortality through the early detection of health risks among high-risk populations such as smokers (Raja et al., 2024; Ali and Ahmed, 2025). Nonetheless, the challenges of overfitting to shallow models, single-center biases, data imbalance, and insufficient multi-population validation will persist and will have to undergo resolutions to guarantee generalizability and clinical trust (Liu et al., 2024; Ali and Ahmed, 2025). The CH route is in the future to integrate wearable sensors to monitor HRV continuously, then there will be multi-center EHR validation, SHAP-based interpretability, and embedding the models into clinical decision support systems to close the gap between research and practice (Raja et al., 2024; Liu et al., 2024).

Shashikant and Chetankumar (2019) developed a machine learning–based predictive framework specifically targeting cardiac arrest among smokers, using Heart Rate Variability (HRV) indices as physiological biomarkers. Their dataset included 1,562 middle-aged individuals, of which 811 were smokers, and 19 HRV parameters were used to represent hemodynamic, time-domain, frequency-domain, and nonlinear cardiac dynamics. The study compared the performance of Logistic Regression (LOR), Decision Tree (DT), and Random Forest (RF) classifiers under an 80:20 train-test split and 10-fold cross-validation. Results showed that RF achieved the highest classification accuracy (93.61%), followed by DT (92.59%) and LOR (88.50%), with RF also producing the best sensitivity, specificity, F1 score, and AUC values. The findings highlight that smoking-related cardiac dysfunction manifests significantly in HRV patterns, and ensemble-based ML models can capture this variability more effectively than linear models. This study aligns with the emerging evidence that non-linear ML architectures outperform traditional approaches in identifying cardiovascular risk among smokers, reinforcing the role of HRV as a predictive indicator for smoking-induced cardiac events.

Siegel et al. (2025) conducted a machine–learning–based analysis to identify patterns of tobacco exposure that contribute to cardiovascular disease (CVD) risk. The study utilized a biomarker dataset containing multiple tobacco-metabolite indicators and applied unsupervised clustering to distinguish smoking exposure phenotypes. Following this, the authors evaluated the predictive performance of several supervised ML models, including Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machine (SVM), and Logistic Regression, to classify individuals at elevated cardiovascular risk. Among the compared models, GBM demonstrated the highest predictive performance, achieving superior accuracy and AUC metrics compared to RF and SVM, while Logistic Regression performed the weakest due to its linear decision boundary. The study highlights that the risk of CVD increases not only with smoking status but also with specific exposure intensity patterns, and further confirms that ensemble ML methods are better suited to capturing non-linear physiological effects of smoking on cardiovascular function. This aligns with other findings emphasizing that non-linear models demonstrate improved performance in predicting smoking-related cardiovascular outcomes.

Alaa et al. (2019) conducted a large-scale prospective study to evaluate the effectiveness of automated machine learning (AutoML) in predicting cardiovascular disease (CVD) risk using data from 423,604 participants of the UK Biobank, where smoking status was included among the clinical and lifestyle predictors. The study implemented the AutoPrognosis framework, which automatically selected and optimized machine learning pipelines and compared its predictive performance to both the Framingham Risk Score and a Cox Proportional Hazards (Cox PH) model. Results demonstrated that AutoPrognosis achieved a higher AUC (0.774; 95% CI: 0.768–0.780) compared to the Framingham score (AUC: 0.724) and Cox PH (AUC: 0.757), indicating improved risk discrimination. The findings reinforce that incorporating smoking-related variables into machine learning–based risk models enhances predictive capability and that non-linear and ensemble-based ML systems outperform conventional statistical models when assessing smoking-associated cardiovascular risk.

Hossain et al. (2024) conducted a cross-sectional study in Bangladesh to develop a machine learning–based framework for predicting cardiovascular disease risk using a combination of clinical, behavioral, and lifestyle variables. The dataset included 391 diagnosed CVD patients and 260 non-CVD individuals, and seven ML models: Logistic Regression, Naïve Bayes, Decision Tree, AdaBoost, Random Forest, Bagging Tree, and Ensemble Learning were trained and evaluated using an 80:20 train-test split. Among all models, the Random Forest classifier demonstrated the highest overall performance, achieving 98.04% accuracy, 96.15% precision, 100% recall, and the highest AUC (0.989), indicating strong discriminative power and robustness against noise and non-linear interactions. In comparison, Logistic Regression showed the lowest performance with 95.42% accuracy due to its limited ability to capture complex nonlinear feature relationships.

The study also incorporated SHAP-based interpretability to rank influential features, showing that high cholesterol, hypertension, irregular heart rhythm, obesity, and smoking behavior were among the strongest predictors of CVD risk. Notably, individuals reporting smoking behavior had a significantly higher likelihood of developing CVD, reinforcing smoking as a major independent risk factor. The overall findings support the effectiveness of ensemble-based models, particularly Random Forest, for identifying high-risk individuals and improving early detection strategies in clinical decision-support systems (Hossain et al., 2024).

Rawat et al. (2023) conducted a comprehensive review highlighting tobacco smoking as a major preventable risk factor for coronary heart disease (CHD), showing that smoking, whether active, passive, or smokeless, contributes to CHD development through mechanisms including endothelial dysfunction, inflammation, thrombosis, and disrupted lipid profiles. The review emphasized that even low levels of smoking significantly increase CHD risk without a safe threshold, with smoking responsible for 1.62 million global CHD deaths and considerable disability-adjusted life years lost. Passive smoking also raises CHD risk by 23-30%, while smokeless tobacco, especially prevalent in Asia, is linked to increased cardiovascular events. Smoking worsens hypertension and cerebral blood flow by impairing nitric oxide production, thus increasing stroke and dementia risks. Despite these dangers, smoking cessation remains the most effective intervention to reduce CHD risk, requiring integrated public health and healthcare system efforts for tobacco control and cessation support (Rawat et al., 2023; Puig-Cotado et al., 2020; Banks et al., 2019; Gallucci et al., 2020; Gupta et al., 2019; Virdis et al., 2010; Joossens et al., 2020; Arnett et al., 2019; Ma et al., 2024).

Wells (1994) carried out a thorough review and risk analysis of the epidemiologic, physiologic, and biochemical evidence linking passive smoking, or environmental tobacco smoke, to ischemic heart disease. The review synthesized data from multiple epidemiologic studies conducted over many years that evaluated risk in never smokers exposed to passive smoke, revealing relative risks ranging approximately from 1.2 to 2.7 for IHD morbidity and mortality, with higher quality studies tending to report greater risk estimates (Wells, 1994). Biologic plausibility for these findings is supported by such mechanisms as increased platelet sensitivity, endothelial damage, oxygen transport impairment due to carbon monoxide, and accelerated atherosclerotic plaque development-effects observed both acutely and in long-term animal and human studies (Wells, 1994). The estimated attributable mortality from passive smoking in the United States in 1985 was about 62,000 ischemic heart disease deaths, substantially higher than earlier estimates because of improvements in risk modeling and additional data sources (Wells, 1994). Though there are some limitations in this study, such as potential confounding factors and publication bias, the review concludes that a well-established causal relationship exists and recommends avoiding exposure to environmental tobacco smoke, particularly for populations at risk of cardiovascular disease (Wells, 1994). This work importantly highlights passive smoking as a major, preventable risk factor for cardiovascular disease.

Glantz and Parmley (1991) provided an extensive review that summarized passive smoking, or environmental tobacco smoke, enhances the risk of coronary heart disease by about 30%, a fact supported by consistent epidemiologic studies and robust biological mechanisms, including endothelial dysfunction, increased platelet aggregation, impaired oxygen transport, and accelerated atherosclerosis due to carcinogenic compounds in smoke. It points out that nonsmokers are particularly sensitive to the exposures, suffering similar acute and chronic cardiovascular impairment as smokers. This work established passive smoking as a major, preventable cardiovascular risk factor causing far more deaths than lung cancer attributed to secondhand smoke and made it the third leading preventable cause of death in the United States after active smoking and alcohol.

Stallones, in 2015, reviewed the extensive and complex literature that concerned the relationship of tobacco smoking to coronary heart disease. The review underlines that serious challenges continue to exist regarding generating consistent conclusions from the different epidemiological studies with diverse methodologies and populations with potential biases. The primary findings indicate that cigarette smokers experience about twice the risk of developing CHD compared to non-smokers; the risk usually increases with the amount smoked, although it decreases with advancing age, though variability among studies is described (Stallones, 2015). Specificity and consistency of the association are further discussed with regard to the fact that smoking is associated with many causes of mortality, though it is strongly implicated in myocardial infarction and CHD. With conflicting data regarding intermediary factors like serum lipids, body weight, and blood pressure, epidemiological evidence suggests a very strong case for causality between smoking and CHD. Stallones emphasizes the need for future research amid recommendations for public health measures to lower the prevalence of smoking because of the huge burden its cardiovascular health carries with it (Stallones, 2015).

References:

1. Raj, S., Alam, S., & Ahmad, T. (2024). Predict the Possibility of Cardiac Arrest in Smokers using Machine Learning Models. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4910651
2. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080

3. Ali, S. N. M., & Ahmed, N. M. (2025). Comparing some Machine Learning Models for Cardiovascular Disease. Journal of Pioneering Medical Science, 14(04), 60–66. https://doi.org/10.47310/jpms2025140408
4. Shashikant, R., & Chetankumar, P. (2019). *Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter.* Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2019.06.002
5. Siegel, M., Agarwal, V., & Khan, T. (2025). *Identifying Patterns of Tobacco Use and Associated Cardiovascular Disease Risk Through Machine Learning Analysis of Urine Biomarkers.* Patterns. https://doi.org/10.1016/j.jacadv.2025.101630
6. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE, 14*(5), e0213653. https://doi.org/10.1371/journal.pone.0213653
7. Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). *Machine learning approach for predicting cardiovascular disease in Bangladesh: Evidence from a cross-sectional study in 2023.* BMC Cardiovascular Disorders, 24(214), 1–28. https://doi.org/10.1186/s12872-024-03883-2
8. Hivre, M. D., Rawat, A., Hasan, M. H., & others. (2023). Smoking and Coronary Heart Disease Impact. Journal of Pharmaceutical Negative Results, 14(S2), 210. https://www.researchgate.net/publication/368282289_Smoking_And_Coronary_Heart_Disease_Impact

9. Wells, A. J. (1994). Passive smoking as a cause of heart disease. Journal of the American College of Cardiology, 24(2), 546-554. https://doi.org/10.1016/0735-1097(94)90315-8
10. Glantz, S. A., & Parmley, W. W. (1991). Passive smoking and heart disease: Epidemiology, physiology, and biochemistry. Circulation, 83(1), 1-12. https://doi.org/10.1161/01.cir.83.1.1
11. Stallones, R. A. (2015). The association between tobacco smoking and coronary heart disease. International Journal of Epidemiology, 44(3), 735-743. https://doi.org/10.1093/ije/dyv124