Evaluating Cardiovascular Disease Prediction Models Across Demographic Subgroups
June 7, 2025
By: Thanh Mai

I. INTRODUCTION

Cardiovascular disease (CVD) remains the leading cause of death worldwide (CDC, 2024), with well-documented disparities in diagnosis, treatment, and outcomes across demographic groups such as sex, race, and age (Niakoeui et al., 2020). These disparities reflect broader structural inequalities that continue to shape healthcare access and quality. As artificial intelligence (AI) and machine learning become increasingly integrated into healthcare, they offer new opportunities to improve disease prediction, efficiency, and personalized care (Armoundas et al., 2024; Yan et a., 2019). However, if not critically evaluated, machine learning models can inadvertently replicate or worsen existing disparities, especially in conditions like CVD where these inequities are already pronounced (Thamman et al., 2023).

This study examines whether a Random Forest classifier trained to predict cardiac events performs consistently across demographic subgroups. We evaluate model accuracy and error rates by sex, race/ethnicity, and age group to assess whether performance disparities exist. Understanding these differences is essential to ensuring that AI applications in healthcare promote equity rather than reinforce bias.

II. METHODS

This study used data from the 2022 Behavioral Risk Factor Surveillance System, a nationwide health survey conducted annually by the U.S. Centers for Disease Control and Prevention (CDC, 2022). A processed version from Kaggle (Pytlak, 2021) was used, containing health and demographic variables relevant to cardiovascular risk. Thirty-nine features were selected based on clinical relevance, including direct indicators (e.g., high blood pressure, diabetes) and indirect factors (e.g., income, education). The target variable was a combined binary indicator of self-reported cardiac events defined as having experienced a heart attack, stroke, or angina.

Initial exploratory data analysis was conducted within each demographic subgroup to examine sample distribution and trends. A Random Forest model was then trained using a balanced class weighting scheme and optimized through grid search with 3-fold cross-validation. The model was evaluated on the test set using AUC, precision, recall, and F1 score. The primary focus of the study was subgroup analysis. Performance metrics were stratified by sex, race/ethnicity, and age to investigate potential disparities in model effectiveness.

III. RESULTS

A. Demographic Summary

A preliminary exploratory data analysis was conducted to examine demographic patterns related to cardiac events. The dataset included sex (e.g., male, female), race/ethnicity (e.g., White non-Hispanic, Hispanic, Black, and others), and age (13 categories from 18 to 99+ years). Males accounted for a greater share of cardiac events (about 57%), and White non-Hispanic individuals were the majority. Hispanic and other minority groups were underrepresented. Cardiac events were strongly associated with age, with older adults comprising most cases.

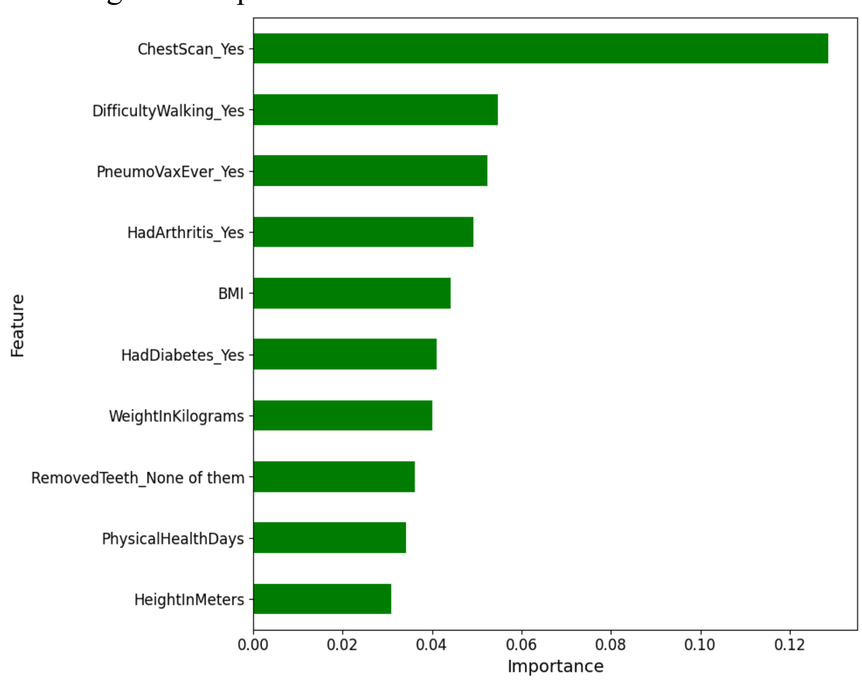
B. Overall Model Performance and Feature Importance

The Random Forest model achieved an ROC AUC value of 0.89 with a 79% accuracy in predicting cardiac events. It showed strong precision and recall for identifying individuals without cardiac events but was less precise though moderately sensitive in detecting those with cardiac events (**Table 1**). The F1 score also reflects this imbalance, indicating that the model is better at ruling out cardiac events than confirming them. The model's top predictors included having had a chest scan, difficulty walking, and pneumococcal vaccination history **(Figure 1)**. Chronic conditions like arthritis and diabetes also ranked highly, along with physical health metrics such as BMI, weight, and height. These features reflect a clinical history and physical status often associated with cardiovascular risk.

Table 1. Overall Model Performance on Cardiac Event Prediction

Class	**Precision**	**Recall**	**F1-Score**	**Support**
No Cardiac Event	0.95	0.80	0.87	71,876
Had Cardiac Event	0.31	0.70	0.43	9,312
Accuracy			0.79	81,188

Figure 1. Top 10 Features in the Random Forest Model



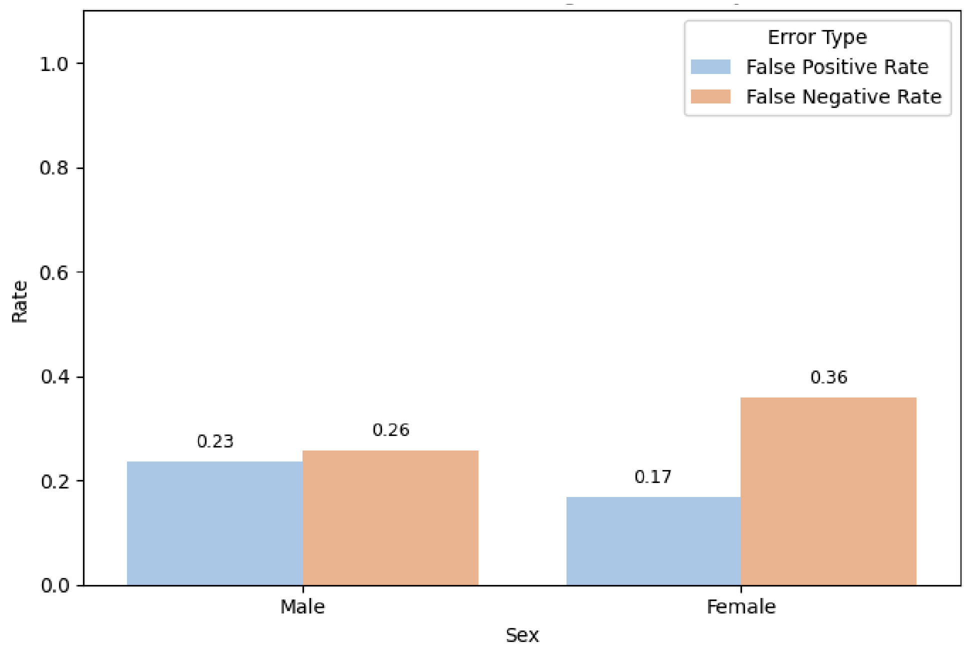
C. Model Performance Across Demographic Subgroups: Sex

The model had similar AUC for both sexes, indicating comparable ability to predict cardiovascular risk. However, recall and precision were both slightly lower for females **(Table 2)**. Additionally, error rate analysis revealed differences between sexes **(Figure 2)**. Males had a false positive rate of 0.23 and a false negative rate of 0.26. Females showed a lower false positive rate (0.17) but a notably higher false negative rate (0.36).

Table 2. Model Performance by Sex on Cardiac Event Prediction

Sex	**AUC**	**Precision**	**Recall**	**F1**	**Count**
Female	0.829	0.290	0.644	0.400	42,107
Male	0.834	0.329	0.746	0.457	39,081

Figure 2. False Positive and False Negative Rates By Sex



D. Model Performance Across Demographic Subgroups: Race/Ethnicity

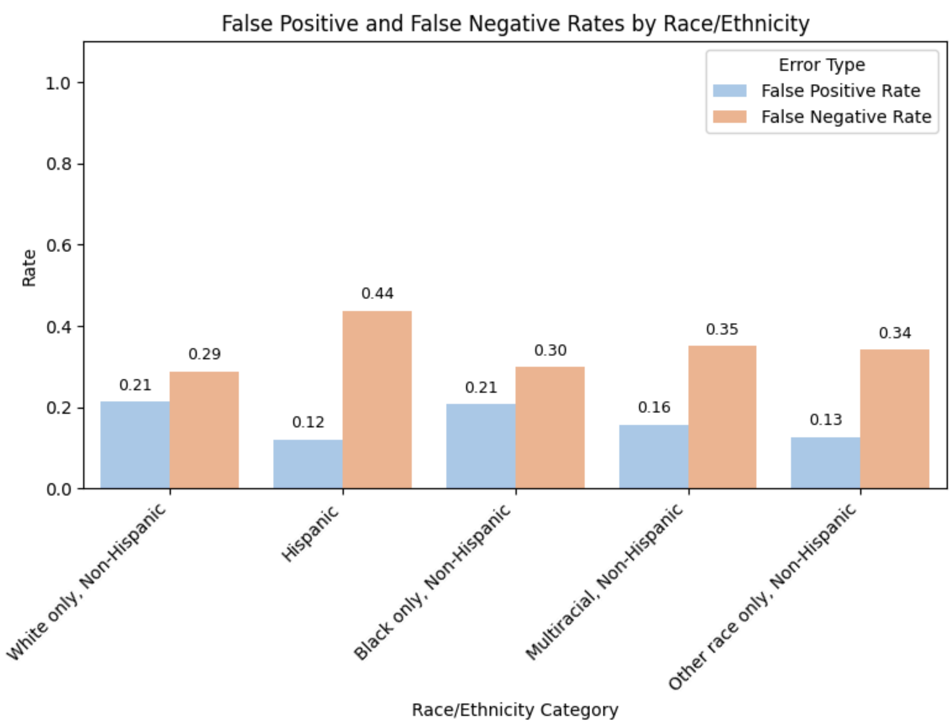
All race/ethnicity groups achieved similar AUC, but Hispanic individuals had the lowest F1 score and precision, suggesting the model may overpredict cardiac events for this group (**Table 3**). White and Black, non-Hispanic individuals performed similarly, but Multiracial and Other race groups had slightly better F1 and precision, though with smaller sample sizes.

False positive and false negative rates varied across racial and ethnic groups. White non-Hispanic individuals had a false positive rate of about 0.21 and a false negative rate of 0.29 (**Figure 3**). Hispanic participants showed a lower false positive rate (0.12) but a much higher false negative rate (0.44), indicating more missed cardiac events. Other groups had false negative rates between 0.30 and 0.35.

Table 3. Model Performance By Race/Ethnicity on Cardiac Event Prediction

Race/Ethnicity	AUC	Precision	Recall	F1	Count
White only, Non-Hispanic	0.831	0.314	0.714	0.436	61,483
Black only, Non-Hispanic	0.823	0.292	0.700	0.412	6,426
Hispanic	0.835	0.269	0.579	0.367	7,425
Other race only, Non-Hispanic	0.865	0.353	0.661	0.460	4,042
Multiracial, Non-Hispanic	0.831	0.369	0.649	0.471	1,812

Figure 3. False Positive and False Negative Rates By Race/Ethnicity



E. Model Performance Across Demographic Subgroups: Age

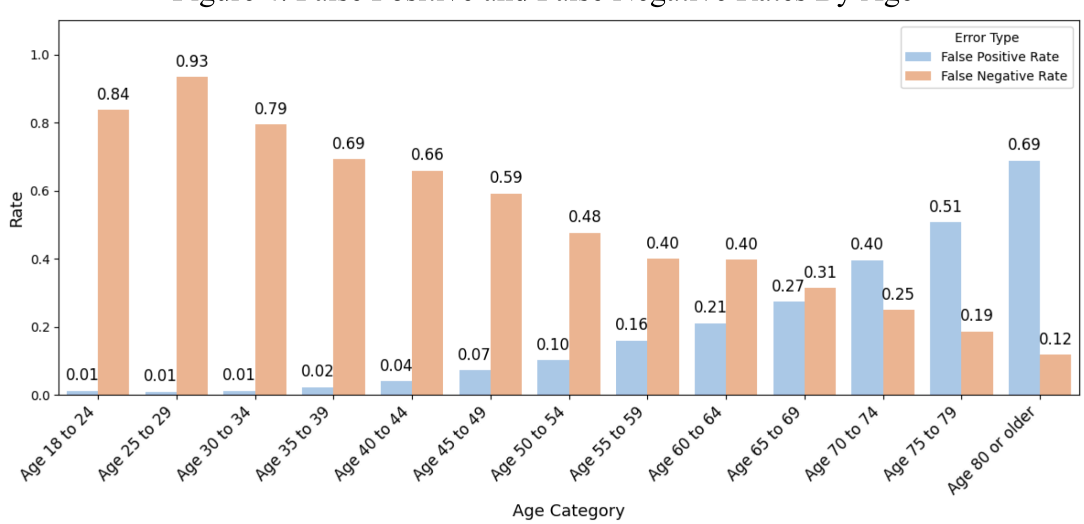
Model performance improved with increasing age **(Table 4)**. While AUC declined slightly for older adults, recall and F1 scores increased, peaking in the 80+ group. Performance was notably poor in younger adults (under 40), with F1 scores below 0.30 and recall as low as 0.07 (age 25–29).

The analysis of error rates across age categories revealed a striking inverse pattern between false negative and false positive rates **(Figure 4)**. Younger age groups, particularly those under 40, exhibited notably high false negative rates, indicating the model frequently missed true cardiac events in these populations. In contrast, false negative rates steadily declined with increasing age. Conversely, false positive rates were low among younger individuals but increased progressively in older age groups, suggesting the model was more likely to incorrectly predict cardiac events as age increased.

Table 4. Model Performance By Age on Cardiac Event Prediction

Age Group	AUC	Precision	Recall	F1	Count
18–24	0.758	0.127	0.186	0.151	4,327
25–29	0.720	0.091	0.067	0.077	3,582
30–34	0.815	0.167	0.176	0.171	4,334
35–39	0.820	0.237	0.323	0.273	5,155
40–44	0.806	0.225	0.358	0.276	5,520
45–49	0.821	0.223	0.398	0.286	5,535
50–54	0.817	0.293	0.530	0.377	6,592
55–59	0.804	0.301	0.603	0.402	7,496
60–64	0.773	0.290	0.606	0.393	8,689
65–69	0.771	0.313	0.692	0.431	9,398
70–74	0.741	0.306	0.748	0.435	8,676
75–79	0.714	0.324	0.806	0.462	5,995
80+	0.697	0.350	0.891	0.503	5,889

Figure 4. False Positive and False Negative Rates By Age



IV. DISCUSSION

A. Overview of Results

The model demonstrated moderate overall performance in predicting cardiac events; however, a clear disparity emerged across the demographic subgroups, raising concerns about machine learning and equity in clinical use. Females had notably lower precision and recall than males, suggesting the model may be less sensitive to detecting cardiac events in women. When stratified by race and ethnicity, Hispanic and multiracial individuals showed higher false negative rates. Age also played a key role. Older adults had higher false positive rates, while younger adults experienced sharply elevated false negative rates. This pattern may reflect the heavier presence of older individuals with documented heart conditions in the dataset, potentially skewing the model's predictive confidence toward that population. These findings suggest the model performs best for groups most represented in the training data, emphasizing the need for careful evaluation before clinical deployment.

B. Limitations & Future Studies

Several limitations should be considered when interpreting the findings of this analysis. As previously noted, the model demonstrated slightly weaker performance for females, particularly in recall and precision. This may reflect broader clinical challenges, as women often present with less typical symptoms of cardiac events, which can complicate diagnosis and potentially affect model sensitivity. Second, the race and ethnicity categories used in the dataset were limited to broad groups (e.g., White non-Hispanic, Black non-Hispanic, Hispanic), omitting key populations such as Asian American, Native Hawaiian, Pacific Islander, and American Indian and Alaskan Native individuals. This restricts the generalizability of subgroup findings and limits a full assessment of model fairness across all racial and ethnic groups. More granular and inclusive demographic classifications would strengthen future analyses. Finally, the model performed less accurately for younger age groups, possibly due to a variety of factors such as healthier conditions or even limited exposure to diagnostic procedures like chest scans, a top predictive feature in the model. Younger adults may have fewer records available, reducing prediction quality. Additionally, the use of thirteen separate age categories may have fragmented the data, decreasing the reliability of subgroup metrics. Collapsing these into broader age bins could improve interpretability.

Another important limitation to note of this study is class imbalance. The dataset contains a much higher proportion of individuals without cardiac events compared to those who experienced one. This imbalance can bias the model toward the majority class, potentially leading to high overall accuracy but poor sensitivity in detecting actual cardiac events, especially in subgroups where the event rate is even lower. Although metrics like recall and AUC help account for this, the underlying skew can still hinder the model's use in the real world, particularly in identifying true positives in marginalized or smaller communities.

V. CONCLUSION

Overall, these results highlight the need for careful evaluation of AI models in healthcare. Without addressing the observed performance disparities, such models risk perpetuating or even worsening existing health inequities. Importantly, the composition of the training data plays a critical role. When certain populations are underrepresented, the model is less likely to learn meaningful patterns for those groups, leading to reduced accuracy and fairness. Thorough subgroup analyses, along with efforts to improve data diversity, are essential to ensure that AI-driven tools deliver equitable care across all patient populations.

REFERENCES

Armoundas, A. A., Narayan, S. M., Arnett, D. K., Spector-Bagdady, K., Bennett, D. A., Celi, L. A., Friedman, P. A., Gollob, M. H., Hall, J. L., Kwitek, A. E., Lett, E., Menon, B. K., Sheehan, K. A., & Al-Zaiti, S. S. (2024). Use of artificial intelligence in improving outcomes in heart disease: A scientific statement from the American Heart Association. Circulation, 149(14). https://doi.org/10.1161/cir.0000000000001201

Centers for Disease Control and Prevention (2022). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/

Centers for Disease Control and Prevention. (2024, October 25). FASTSTATS - leading causes of death. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Niakouei, A., Tehrani, M., & Fulton, L. (2020). Health Disparities and Cardiovascular Disease. Healthcare (Basel, Switzerland), 8(1), 65. https://doi.org/10.3390/healthcare8010065

Pytlak, K. (2023, October 12). Indicators of heart disease (2022 update). Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Thamman, R., Yong, C. M., Tran, A. H., Tobb, K., & Brandt, E. J. (2023). Role of artificial intelligence in Cardiovascular Health Disparities. JACC: Advances, 2(7), 100578. https://doi.org/10.1016/j.jacadv.2023.100578

Yan, Y., Zhang, J. W., Zang, G. Y., & Pu, J. (2019). The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine?. Journal of geriatric cardiology : JGC, 16(8), 585–591. https://doi.org/10.11909/j.issn.1671-5411.2019.08.010