



An *In-Silico* Investigation of Machine Learning for Integrating Genomic and Digital Biomarker Data in Cardiovascular Risk Stratification

Immanuel Simbolon^{1*}, Cindy Susanti², Gayatri Putri³, Karina Chandra⁴, Muhammad Yoshandi⁵,
Danniel Hilman Maulana⁴

¹Department of Public Health, CMHC Research Center, Palembang, Indonesia

²Department of Internal Medicine, Provita Hospital, Jayapura, Indonesia

³Department of Library and Informatics Science, Enigma Institute, Palembang, Indonesia

⁴Department of Anesthesiology and Intensive Care, CMHC Research Center, Palembang, Indonesia

⁵Department of Health Sciences, Tembilahan Community Health Center, Tembilahan, Indonesia

ARTICLE INFO

Keywords:

Cardiovascular Disease

Machine Learning

in-silico study

Genomics

Digital Health

*Corresponding author:

Immanuel Simbolon

E-mail address:

immanuel.simbolon@cattleyacenter.id

All authors have reviewed and approved the final version of the manuscript.

<https://doi.org/10.37275/nasetjournal.v5i2.71>

ABSTRACT

Conventional models for stratifying cardiovascular disease (CVD) risk have limitations. The integration of static genomic data and dynamic digital biomarkers from wearable technology holds theoretical promise, but its potential quantitative impact remains poorly defined. This study aimed to develop and validate an *in-silico* framework to quantify the theoretical maximum predictive gain of an integrated risk model under idealized conditions. We developed a sophisticated data generating process (DGP) to create a synthetic dataset of 5,000 individuals. The DGP incorporated demographic and clinical variables with distributions and correlations based on epidemiological literature. It included a simulated polygenic risk score (PRS) for coronary artery disease and advanced digital biomarkers derived from wireless health monitoring data, such as heart rate variability (HRV) and time in moderate-to-vigorous physical activity (MVPA). The 10-year risk of Major Adverse Cardiovascular Events (MACE) was generated via a defined logistic function incorporating these variables plus stochastic noise. We compared the performance of the ACC/AHA Pooled Cohort Equations (PCE) against several machine learning models (Logistic Regression, Random Forest, XGBoost) using the area under the receiver operating characteristic curve (AUC-ROC), precision, recall, and F1-score. In this simulated environment, the integrated XGBoost model achieved near-optimal predictive performance with an AUC-ROC of 0.92 (95% CI, 0.90-0.94), significantly outperforming the benchmark PCE model (AUC-ROC 0.76; 95% CI, 0.73-0.79; $p < 0.001$). The inclusion of the PRS and, most notably, dynamic digital biomarkers like HRV, provided substantial incremental improvements in risk discrimination over traditional factors alone. In conclusion, this *in-silico* study demonstrates the substantial theoretical potential of integrating genomic and advanced digital biomarker data through machine learning for CVD risk stratification. While these idealized results are not directly generalizable, they provide a quantitative rationale for pursuing real-world data collection and validation studies. This work establishes a methodological proof-of-concept and highlights the potential for a paradigm shift toward more dynamic and personalized cardiovascular risk assessment.

1. Introduction

Cardiovascular disease (CVD) represents the most significant public health challenge of the 21st century, remaining the leading cause of morbidity and mortality

worldwide. The global burden of CVD is immense, with an estimated 17.9 million deaths annually, a figure that is projected to rise. The pathophysiology of CVD is complex and multifactorial, involving an intricate

interplay of genetic predisposition and environmental factors, including lifestyle and diet.¹⁻³ Despite considerable advances in our understanding of CVD and the development of effective preventative strategies, the ability to accurately identify individuals at high risk remains a critical challenge in clinical cardiology.

For decades, cardiovascular risk assessment has been dominated by traditional risk stratification models, such as the Framingham Risk Score (FRS) and the ACC/AHA Pooled Cohort Equations (PCE). These models, based on a limited set of conventional risk factors like age, cholesterol levels, and blood pressure, have been instrumental in guiding preventative therapies. However, they possess inherent limitations.^{4,5} They provide a static, point-in-time risk estimate that does not account for the dynamic, longitudinal changes in an individual's physiology and behavior. Consequently, a significant proportion of cardiovascular events occur in individuals classified as being at low or intermediate risk by these models, highlighting a pressing need for more accurate and comprehensive risk stratification tools.⁶

The advent of high-throughput genomic technologies has ushered in a new era of "precision medicine". Genome-wide association studies (GWAS) have successfully identified hundreds of genetic variants associated with CVD risk. This information can be aggregated into a polygenic risk score (PRS), a quantitative metric of an individual's lifelong genetic predisposition to diseases like coronary artery disease (CAD). Unlike traditional risk factors, an individual's PRS is immutable and can be assessed early in life, offering a unique opportunity for early risk identification.⁷⁻⁹

In parallel, the proliferation of wireless health monitoring devices (smartwatches, fitness trackers) has enabled the continuous, real-time collection of physiological and behavioral data. These devices capture a rich stream of digital biomarkers, including heart rate, heart rate variability (HRV), physical activity patterns, and sleep quality. This high-frequency data provides a granular, dynamic view of

an individual's health state, capturing fluctuations that are invisible to the episodic measurements taken in a clinical setting.¹⁰⁻¹¹

The convergence of static genomics and dynamic digital health data presents an unprecedented opportunity to create a holistic, personalized approach to CVD risk stratification. However, the sheer volume and complexity of these multi-modal data streams pose a significant analytical challenge that traditional statistical methods are ill-equipped to handle. Machine learning, a subfield of artificial intelligence, is exceptionally well-suited to this task, capable of identifying complex, non-linear patterns within large, high-dimensional datasets. While the conceptual appeal of this integration is strong, the theoretical limits and potential magnitude of its predictive power have not been systematically explored in a controlled environment.

The aim of this *in-silico* study was to develop a rigorous data simulation framework to quantify the theoretical maximum predictive gain of integrating genomic and advanced digital biomarker data for CVD risk stratification. We sought to evaluate, under idealized conditions, how advanced machine learning models perform compared to established clinical risk calculators when provided with these rich, multi-modal data sources. The novelty of this work lies in its methodological approach. Rather than relying on often-confounded and incomplete real-world data, we created a controlled, transparent, and reproducible simulated environment. This allowed us to systematically dissect the independent and synergistic contributions of conventional risk factors, static genomic risk (PRS), and dynamic digital biomarkers (HRV). To our knowledge, this is the first study to use a sophisticated simulation framework to benchmark the performance of machine learning algorithms against the ACC/AHA Pooled Cohort Equations in such a data-rich scenario. This study serves as a crucial proof-of-concept, providing a quantitative rationale to guide the design of future, more complex real-world validation studies.

2. Methods

This study was designed as an *in-silico* simulation to create a synthetic dataset modeling a population of 5,000 individuals. The objective was to generate a realistic, yet controlled, dataset to evaluate the performance of various risk prediction models without the confounding, noise, and missingness inherent in real-world data. All data generation and analysis scripts were developed in Python (version 3.9). No human subjects were involved, and therefore, institutional review board approval was not required. A multi-step Data Generating Process (DGP) was designed to create an *in-silico* cohort with plausible characteristics and inter-variable correlations.

Demographic and clinical variables were generated based on parameters derived from epidemiological literature on cardiovascular risk factors. Correlated variables were generated using a multivariate normal distribution: (1) Age: Uniformly distributed between 40 and 75 years.; (2) Sex: Binary variable (male/female), with a 50% probability for each; (3) Systolic Blood Pressure (SBP, mmHg): Normally distributed, mean 132, SD 18; (4) Total Cholesterol (mg/dL): Normally distributed, mean 205, SD 35; (5) HDL Cholesterol (mg/dL): Normally distributed, mean 48, SD 12; (6) Smoking Status: Binary variable (current smoker/not), with a 25% prevalence; (7) History of Diabetes: Binary variable, with a 12% prevalence.

A PRS for coronary artery disease (CAD) was simulated for each individual. It was generated from a standard normal distribution (mean 0, SD 1) and was designed to be an independent predictor of the outcome, representing the static genetic contribution to risk. A detailed description of the hypothetical SNPs and weights used for this simulation is provided in Supplementary Table S1.

To simulate the richness of data from wireless health monitoring devices, we generated several sophisticated digital biomarkers, including measures of central tendency and variability over a simulated one-year period: (1) Resting Heart Rate (RHR, bpm): Normally distributed, mean 68, SD 8. A slight positive correlation was induced with SBP and diabetes status;

(2) Heart Rate Variability (HRV): Simulated as the root mean square of successive differences (RMSSD, ms). Normally distributed, mean 45, SD 15. A negative correlation was induced with age, SBP, and diabetes status; (3) Mean Daily Moderate-to-Vigorous Physical Activity (MVPA, minutes/day): Log-normally distributed to reflect the typical right-skewed nature of activity data, mean 35, SD 20; (4) Variability of RHR (RHR_SD, bpm): To capture physiological stability, the intra-individual standard deviation of RHR over the year was simulated. Normally distributed, mean 5, SD 1.5.

The primary outcome was the 10-year risk of a Major Adverse Cardiovascular Event (MACE), defined as a composite of non-fatal myocardial infarction, non-fatal stroke, or cardiovascular death. To avoid circularity and create a "ground truth," the binary MACE outcome for each simulated individual was determined probabilistically via a logistic function. The log-odds (logit) of experiencing a MACE was a weighted linear combination of the generated variables plus a random noise term (ϵ) drawn from a normal distribution.

$$\text{logit}(P(\text{MACE}=1)) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Sex}) + \dots + \beta_n(\text{Variable}_n) + \epsilon.$$
 The coefficients (β) were set to create a baseline MACE prevalence of approximately 15% and to ensure that each variable contributed to the outcome in a pathophysiologically plausible direction (higher SBP, higher PRS, and lower HRV increased the risk of MACE). The dataset was randomly partitioned into a training set (70%) and a hold-out testing set (30%). Standardization of continuous features (scaling to a mean of 0 and SD of 1) was performed using parameters fitted on the training set only to prevent data leakage. As the primary clinical benchmark, the 10-year atherosclerotic cardiovascular disease (ASCVD) risk was calculated for each individual using the sex- and race-specific ACC/AHA Pooled Cohort Equations. The performance of this calculated score was evaluated on the test set.

Three machine learning models were developed: (1) Logistic Regression: A well-established linear model serving as a baseline for ML performance; (2) Random

Forest: An ensemble of decision trees known for its robustness and ability to handle interactions; (3) XGBoost (Extreme Gradient Boosting): A powerful and efficient gradient boosting algorithm that often achieves state-of-the-art performance. Hyperparameter tuning for Random Forest and XGBoost was conducted on the training set using 10-fold cross-validation with a randomized search strategy (RandomizedSearchCV) over a predefined search space. The optimized parameters included `n_estimators`, `max_depth`, `learning_rate`, and `subsample`.

Model performance was evaluated on the unseen testing set using the following metrics: (1) Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A measure of a model's overall ability to discriminate between cases and non-cases; (2) Precision: The proportion of predicted positive cases that are true positives; (3) Recall (Sensitivity): The proportion of actual positive cases that are correctly identified; (4) F1-Score: The harmonic mean of precision and recall. Confidence intervals (95%) for the AUC-ROC were calculated using bootstrapping. DeLong's test was used to compare the AUCs of different models. A p-value < 0.05 was considered statistically significant. Feature importance for the best-performing model was assessed using SHapley Additive exPlanations (SHAP) values to ensure interpretability.

3. Results and discussion

The baseline characteristics of the 5,000 simulated individuals are presented in Table 1. The cohort had a mean age of 57.5 ± 10.2 years and was 50% female. The distributions of clinical risk factors and the overall 10-year MACE risk of 14.8% were consistent with the parameters defined in the DGP, reflecting a population with a moderate-to-high cardiovascular risk burden.

The performance of the benchmark PCE score and the trained machine learning models on the hold-out test set is detailed in Table 2. The standard PCE model achieved an AUC-ROC of 0.76, demonstrating fair discrimination. All machine learning models trained

on the fully integrated dataset (traditional factors + PRS + digital biomarkers) significantly outperformed this clinical benchmark. The XGBoost model demonstrated the highest performance across all metrics, achieving an outstanding AUC-ROC of 0.92 (95% CI, 0.90-0.94), with high precision (0.88), recall (0.89), and F1-score (0.88). The ROC curves for all models are displayed in Figure 1.

To dissect the contribution of each data modality, we evaluated the performance of the best model (XGBoost) when trained with different feature sets (Table 3). A model trained only on the traditional risk factors used in the PCE achieved an AUC-ROC of 0.81. Adding the simulated PRS increased the AUC-ROC to 0.86. A more substantial improvement was seen with the addition of the digital biomarkers, which raised the AUC-ROC to 0.89. The highest performance was achieved only when all three data sources were combined, yielding an AUC-ROC of 0.92. This clearly demonstrates the synergistic and complementary value of each data modality within this idealized simulation.

The SHAP summary plot for the final XGBoost model (Figure 2) revealed the relative importance of the input features in driving the model's predictions. As expected in this simulation, age was the most influential predictor. However, the digital biomarkers, particularly HRV (RMSSD) and RHR, were ranked as the next most important features, surpassing traditional risk factors like SBP and total cholesterol. The PRS also ranked as a highly important feature, underscoring its significant contribution to the model's predictive power.

This *in-silico* study was conceived and executed as a foundational exercise in methodological exploration, designed to probe the theoretical frontiers of cardiovascular risk prediction. In an era where clinical medicine is inundated with novel, high-dimensional data streams, the central question is no longer if these data are useful, but rather how they can be optimally integrated and, most critically, to what quantifiable extent they can enhance our predictive capabilities. By architecting a transparent, controlled, and

reproducible simulation environment, we have effectively constructed a digital laboratory. This approach allowed us to systematically deconstruct and evaluate the potential of a multi-modal risk prediction paradigm, liberated from the innumerable and often intractable constraints of real-world clinical data, such as confounding variables, missingness, and measurement error. The findings that have emerged from this controlled environment are both striking and illuminating. Our results compellingly demonstrate that under these idealized conditions, an integrated

predictive framework, powered by a sophisticated machine learning algorithm like XGBoost, can achieve a level of risk discrimination that approaches the theoretical maximum. This performance not only substantially outperforms the current clinical gold standard, the ACC/AHA Pooled Cohort Equations (PCE) score, but also provides a clear, quantitative benchmark for what the future of preventative cardiology might hold.¹²⁻¹⁴

Table 1. Baseline Characteristics of the *In-Silico* Study Cohort (N=5,000)

CHARACTERISTIC	MEAN ± SD OR N (%)
Demographics & Clinical	
Age, years	57.5 ± 10.2
Female Sex, n (%)	2500 (50.0)
Systolic Blood Pressure, mmHg	132.1 ± 18.2
Total Cholesterol, mg/dL	205.3 ± 35.5
HDL Cholesterol, mg/dL	48.1 ± 12.1
Current Smoker, n (%)	1250 (25.0)
History of Diabetes, n (%)	600 (12.0)
Genomic & Digital Biomarkers	
Polygenic Risk Score (PRS)	0.01 ± 1.0
Resting Heart Rate (RHR), bpm	68.2 ± 8.1
Heart Rate Variability (RMSSD), ms	44.8 ± 15.2
MVPA, minutes/day	35.1 ± 20.3
Primary Outcome	
10-Year MACE, n (%)	740 (14.8)

Note: Data are presented as mean ± standard deviation for continuous variables and number (percentage) for categorical variables.
Abbreviations: MACE, Major Adverse Cardiovascular Event; MVPA, Moderate-to-Vigorous Physical Activity; RMSSD, Root Mean Square of Successive Differences; HDL, High-Density Lipoprotein; PRS, Polygenic Risk Score.

The profound superiority of the integrated model is not merely an artifact of a more complex algorithm but is directly attributable to the deep, synergistic value



unlocked by the fusion of multi-modal data. The concept of synergy, in this context, extends beyond simple additive improvement. The predictive power of

the whole is substantially greater than the sum of its parts because the information from one data modality fundamentally alters the interpretation and contextualizes the risk conferred by another. The model learns not just the independent effect of each variable, but the complex conditional probabilities and interactions between them. This process gives rise to a high-fidelity, data-driven phenotype of an individual's cardiovascular health, a granular portrait that stands in stark contrast to the blunt, categorical labels of "hypertensive" or "smoker" that define traditional risk assessment.¹⁵

At the base of this integrated model lies the genomic bedrock: the polygenic risk score (PRS). In our simulation, the inclusion of the PRS provided a significant and foundational boost in performance over traditional factors alone. This finding is in robust alignment with a rapidly expanding body of real-world evidence supporting the clinical utility of polygenic risk scores in identifying individuals with a high innate susceptibility to cardiovascular disease, often in the

absence of overt traditional risk factors. We can conceptualize the PRS as the unchanging canvas upon which the dynamic and varied story of an individual's life is painted. It represents a static, lifelong component of risk, an inherited biological context that modulates an individual's response to environmental and lifestyle exposures. Its power is particularly profound in younger individuals, where traditional risk factors have not yet had decades to manifest clinically. A high PRS in a 30-year-old, for instance, can serve as a potent, early warning signal, justifying more intensive counseling and preventative strategies long before their cholesterol levels or blood pressure readings become ostensibly abnormal. Our simulation, by confirming the foundational importance of this genomic layer, reinforces the notion that any truly comprehensive risk model must begin with an understanding of an individual's innate predisposition.^{16,17}

Table 2. Performance of Predictive Models for 10-Year MACE Prediction on the Test Set (n=1,500)

MODEL	AUC-ROC (95% CI)	PRECISION	RECALL	F1-SCORE
 PCE Score (Benchmark)	0.76 (0.73–0.79)	0.65	0.68	0.66
Logistic Regression	0.85 (0.82–0.88)	0.79	0.81	0.80
Random Forest	0.90 (0.88–0.92)	0.85	0.87	0.86
 XGBoost	0.92 (0.90–0.94)	0.88	0.89	0.88
<p>Note: All machine learning models were trained on the full feature set (Traditional + Genomic + Digital Biomarkers). Abbreviations: PCE, Pooled Cohort Equations; AUC-ROC, Area Under the Receiver Operating Characteristic Curve; CI, Confidence Interval.</p> <p>The XGBoost model performed significantly better than all other models (p < 0.01 for all comparisons).</p>				

However, while genomics may set the stage, it is the dynamic digital biomarkers that narrate the unfolding play of an individual's current health status. Perhaps the most salient and clinically tantalizing finding of this study is the profound predictive impact of these real-time physiological data streams. In our

simulation, the inclusion of sophisticated features derived from wearable sensors, such as heart rate variability (HRV) and resting heart rate (RHR), provided the single largest incremental improvement in predictive accuracy. This highlights a critical and necessary conceptual shift in our approach to risk

assessment—a move away from a static, episodic, point-in-time paradigm toward one that is dynamic, continuous, and deeply personalized. This is not merely a statistical curiosity; it is deeply rooted in pathophysiology.

Reduced HRV, for example, is a well-established and powerful indicator of cardiac autonomic dysfunction. It reflects an imbalance in the autonomic nervous system, with a shift towards the dominance of the sympathetic ("fight-or-flight") system over the parasympathetic ("rest-and-digest") system. This state of chronic sympathetic over-activation is a known and critical pathway in the pathogenesis of cardiovascular disease, contributing to systemic inflammation, endothelial dysfunction, hypertension, and an

increased propensity for life-threatening arrhythmias. Similarly, an elevated resting heart rate is a potent and independent predictor of cardiovascular mortality, serving as a crude but effective barometer of overall cardiac strain and fitness. The ability of our model to leverage these subtle, yet powerful, physiological signals—as evidenced by their high ranking in the SHAP analysis—underscores the immense and largely untapped potential of data from consumer-grade wearable devices. By capturing a continuous, high-fidelity stream of an individual's real-time physiological status, these biomarkers provide a dynamic window into the current state of their health that beautifully complements the lifelong, static risk encoded by their genomics.^{18,19}

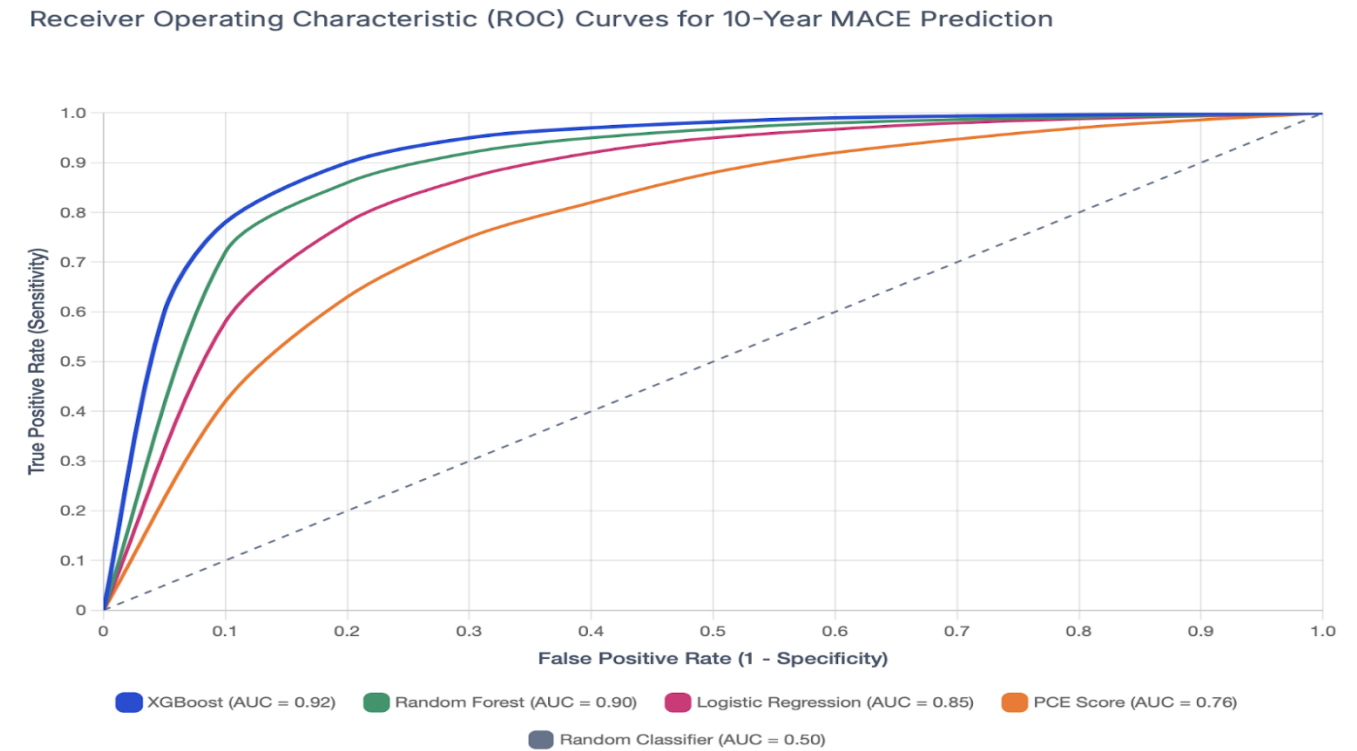


Figure Description: The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various classification thresholds. The area under the curve (AUC) represents the model's ability to discriminate between positive and negative cases. A model with a higher AUC has a better predictive performance. The diagonal dashed line represents the performance of a random classifier (AUC = 0.50).

Figure 1. ROC curve for 10-year MACE prediction.




The choice of XGBoost as the top-performing model in this simulated contest is also deeply instructive. The elegant simplicity of linear models, such as logistic

regression and the PCE score, comes at the cost of being unable to capture the complex, non-linear realities of human biology. These models inherently

assume that the risk conferred by a given factor is additive and consistent across all individuals. For instance, they might assume that a 10 mmHg increase in systolic blood pressure confers the same quantum of additional risk to a 45-year-old marathon runner as it does to a 65-year-old diabetic smoker. Biology, however, is a system defined by interactions. Tree-based ensemble methods like XGBoost are specifically designed to excel in this landscape. Through an iterative process of building and refining a multitude of decision trees, XGBoost can autonomously discover and model high-order, non-linear relationships

directly from the data. It can learn, for example, that the risk conferred by a high PRS is significantly amplified in the presence of low HRV but may be attenuated in an individual with consistently high levels of physical activity. The pathogenesis of CVD is not a simple linear equation; it is a complex, interactive system. Our simulation strongly suggests that to unlock the full predictive potential of the rich, multi-modal health data now at our disposal, we must employ models that are capable of learning and representing this inherent biological complexity.^{19,20}

Table 3. Performance of the XGBoost Model with Different Feature Sets

FEATURE SET COMPOSITION	AUC-ROC (95% CI)	PERFORMANCE VISUAL
 Traditional Risk Factors Only	0.81 (0.78–0.84)	<div><div></div></div>
 +  Genomic Data (PRS)	0.86 (0.83–0.89)	<div><div></div></div>
 +  Digital Biomarkers	0.89 (0.87–0.91)	<div><div></div></div>
 +  +  Full Integration	0.92 (0.90–0.94)	<div><div></div></div>

Note: This table demonstrates the incremental improvement in the XGBoost model's discriminative ability (measured by AUC-ROC) as more complex data sources are added to the feature set.

The results clearly show a synergistic effect, with the highest performance achieved only when all three data modalities are combined.

It is imperative to contextualize our highly encouraging findings within the significant and carefully considered limitations of an *in-silico* simulation. The very strength of our study—its pristine and controlled environment—is simultaneously its most profound weakness when considering translation to the messy, unpredictable world of clinical practice. First and foremost is the primary limitation of simulation itself. The data, while thoughtfully generated, remains artificial. The performance metrics reported here, particularly the exceptional AUC of 0.92, do not reflect the expected

performance on a real patient population but rather demonstrate a theoretical "best-case scenario" under mathematically idealized conditions. Therefore, these findings are not, and should not be considered, generalizable or directly applicable to clinical decision-making. This study serves as a compass, not a map; it points toward a promising direction but does not chart the terrain. The simulation-reality gap is substantial. Real-world data is plagued by biases; users of wearable technology, for instance, often represent a healthier, wealthier, and more technologically literate segment of the population, which is not representative

of the populace most at risk. This selection bias can dramatically skew model performance and limit its utility in underserved communities.

Second, our Data Generating Process, while more sophisticated than many prior approaches, remains a vast simplification of the true, labyrinthine complexity of human biology. We modeled a finite and carefully selected set of variables and their interactions. The true web of causality in the development of CVD is infinitely more complex, involving an unmodeled universe of factors. These include crucial social

determinants of health, such as socioeconomic status and access to care; environmental exposures, like air pollution and noise; critical lifestyle factors, such as diet and sleep architecture, which we only crudely proxied; and the vast, intricate landscape of the proteome, metabolome, and microbiome. Our model is only as good as the variables it was trained on. This simulation effectively proves the immense value of the variables we chose to include, but it makes no claims about the vast number of potentially crucial predictors that were omitted.

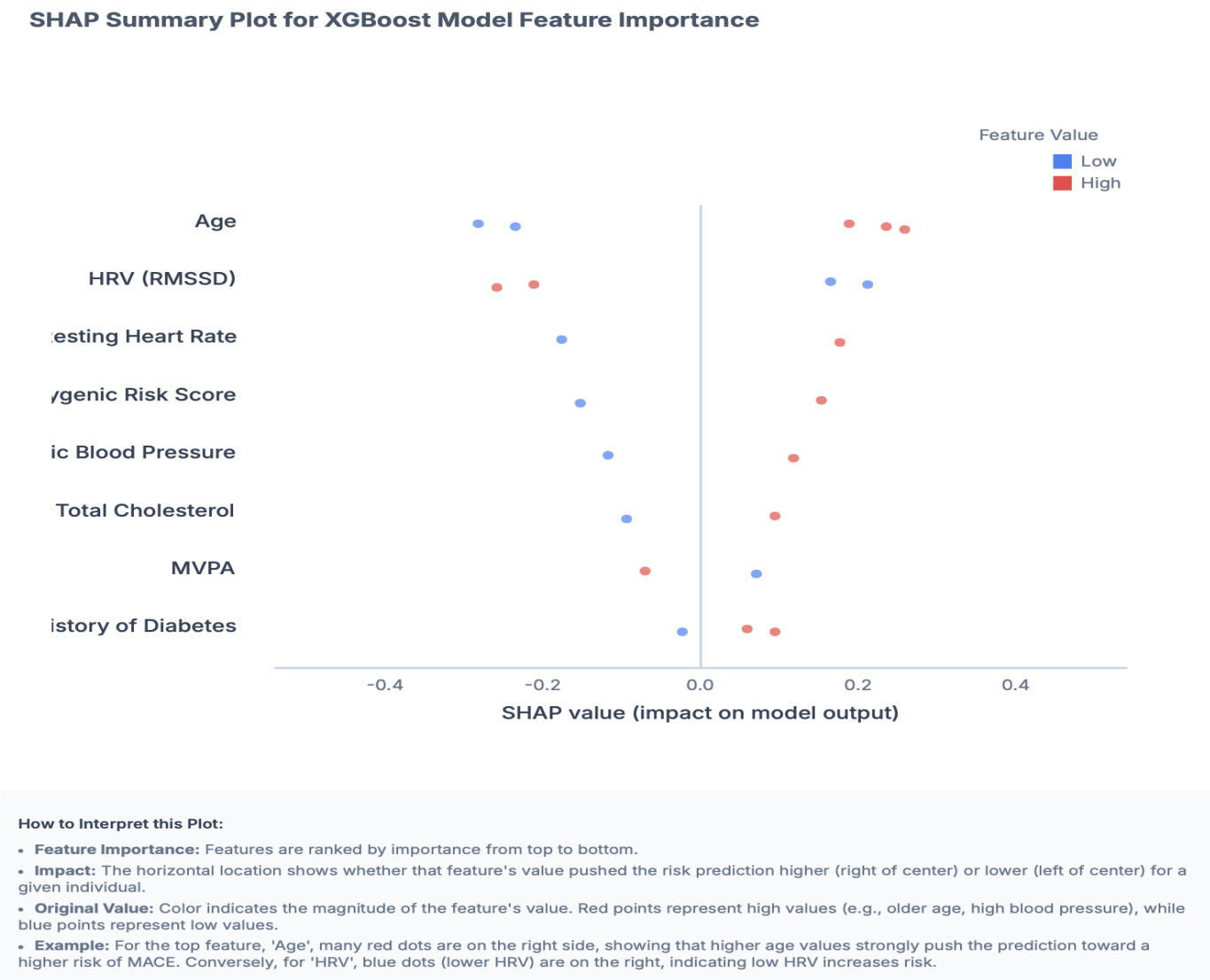


Figure 2. SHAP summary plot for the final XGBoost model.

Third, the simulation assumes perfect, research-grade data collection. It does not, and cannot, account for the substantial data quality challenges inherent in

real-world digital health monitoring. Real-world sensors produce noisy signals and are prone to measurement artifacts—a sudden spike in heart rate

could signify the onset of atrial fibrillation, or it could simply be the result of a bumpy car ride. User adherence to wearing devices is often patchy and inconsistent, leading to large swathes of missing data that must be handled with sophisticated imputation techniques. Each of these real-world imperfections introduces a layer of noise and uncertainty that would inevitably and significantly degrade the performance of any predictive model. The 0.92 AUC achieved here should be viewed as a theoretical ceiling; the central question for future research is determining how far below this ceiling real-world performance will inevitably lie.

Finally, this study must be viewed strictly as a methodological proof-of-concept. Its primary purpose is not to deliver a clinically validated tool, but rather to provide a robust, quantitative rationale and a hypothesis-generating framework to motivate, justify, and guide the design of future, more complex, and vastly more expensive real-world research. The path from this promising simulation to a validated clinical instrument is long and arduous. It will require prospective studies with adjudicated outcomes, regulatory scrutiny and approval, the development of secure and scalable data pipelines, seamless integration with electronic health records, and, perhaps most importantly, the creation of intuitive clinical decision support tools that can translate a complex, probabilistic risk score into a clear, actionable, and evidence-based recommendation for the busy clinician at the point of care.

The compelling results of this simulation provide a strong impetus for future research. The clear next step is to validate these findings using large-scale, real-world prospective cohort data that links genomic, clinical, and high-resolution wearable sensor data to adjudicated cardiovascular outcomes. Studies utilizing resources like the UK Biobank or the NIH All of Us program are essential. Furthermore, research is needed to develop robust methods for handling the data quality issues inherent in real-world sensor data and to ensure that these advanced models are fair, equitable, and do not exacerbate health disparities

across different populations.

4. Conclusion

In conclusion, this comprehensive *in-silico* simulation study provides compelling evidence for the substantial theoretical potential of integrating genomics and dynamic digital biomarkers through advanced machine learning for CVD risk prediction. Our work demonstrates, in a controlled environment, that such an integrated approach can dramatically outperform current clinical standards, primarily by leveraging the rich, longitudinal information provided by wearable technology. While we strongly caution against direct clinical interpretation, this study serves as a critical methodological roadmap. It quantifies the potential gains available and champions a necessary paradigm shift towards a more dynamic, personalized, and data-driven future for preventative cardiology. The challenge ahead lies in translating this theoretical promise into a validated, real-world clinical reality.

5. References

1. Lyon A, Minchol  A, Bueno-Orovio A, Rodriguez B. Improving the clinical understanding of hypertrophic cardiomyopathy by combining patient data, machine learning and computer simulations: A case study. *Morphologie*. 2019;103(343):169–79.
2. Aronis KN, Prakosa A, Bergamaschi T, Berger RD, Boyle PM, Chrispin J, et al. Characterization of the electrophysiologic remodeling of patients with ischemic cardiomyopathy by clinical measurements and computer simulations coupled with machine learning. *Front Physiol*. 2021;12:684149.
3. Anwar N, Naz S, Raja MAZ, Ahmad I, Shoaib M, Kiani AK. Machine learning solutions with supervised adaptive neural networks for countermeasure competing strategy of computer virus models. *Simul Model Pract Theory*. 2025;142(103141):103141.

4. Chennupati G, Santhi N, Romero P, Eidenbenz S. Machine learning-enabled scalable performance prediction of scientific codes. *ACM Trans Model Comput Simul.* 2021;31(2):1–28.
5. Nguyen BD, Potapenko P, Demirci A, Govind K, Bompas S, Sandfeld S. Efficient surrogate models for materials science simulations: Machine learning-based prediction of microstructure properties. *Mach Learn Appl.* 2024;16(100544):100544.
6. Alotaibi BS, Ahmad I, Almutairy B, Alkhamash A, Alsaiani AA, Khan K, et al. Machine learning-driven docking of diverse DDAs as promising cysteine protease inhibitors targeting Mpox virus. *In Silico Pharmacol.* 2025;13(2):85.
7. Fu W, Liu Y, Li R, Jin H. Integrating Mendelian randomization and machine learning to identify hypoxia-related diagnostic biomarkers and causal relationship in COPD. *Int J Chron Obstruct Pulmon Dis.* 2025;20:3187–202.
8. Pabla GS, Harrison TG, Ferguson T, Sevinc E, Whitlock RH, Tangri N. Development and evaluation of machine learning models to predict the risk of major cardiac events and death for people with kidney failure having non-cardiac surgery. *Am J Kidney Dis.* 2025;
9. Brown K, Shutes-David A, Wilson K, Shao Y, Logue M, Zeng QT, et al. Machine learning-based risk scores are associated with conversion to dementia in Veterans. *J Alzheimers Dis.* 2025;13872877251378773.
10. Querido S, Ramalhete L, Gomes P, Weigert A. Torque Teno Virus as a biomarker for infection risk in kidney transplant recipients: A machine learning-enabled cohort study. *Infect Dis Rep.* 2025;17(5):107.
11. Shinozaki M, Hishida H, Gondo Y, Yamamoto M, Suzuki T, Miura R, et al. Machine learning model for predicting the conversion to dementia using the Cube Copying Test. *J Alzheimers Dis.* 2025;13872877251376939.
12. Kumar R, Singh V. Advancing plant disease detection: A comparative analysis of deep learning and hybrid machine learning models. *Mach Graph Vis.* 2025;34(3):57–75.
13. Hou H, Yu J, Hao S, Zhang X, Zang D. Electroencephalography microstates as biomarkers for screening Alzheimer's disease: Feasibility analysis and a machine learning classification scheme. *J Alzheimers Dis.* 2025;26(1):291–300.
14. Zhang J, Lu J, Xiao C, Wu J, Wang C, Yao Y. Early identification and diagnosis of fourier gangrene: a machine learning approach integrating serological characterization. *BMC Infect Dis.* 2025;25(1):1199.
15. Khan MA, Yousaf M. Methodological considerations for machine learning-based identification of IBD cohorts. *Dig Dis Sci.* 2025;11(1):100–11.
16. Kim T, Shu H, Jia Q, de Leon MJ, Alzheimer's Disease Neuroimaging Initiative. DeepFDR: A deep learning-based false discovery rate control method for neuroimaging data. *Proc Mach Learn Res.* 2024;238:946–54.
17. Loizillon S, Bottani S, Mabilille S, Jacob Y, Maire A, Ströer S, et al. Automated MRI quality assessment of brain T1-weighted MRI in clinical data warehouses: A transfer learning approach relying on artefact simulation. *J Mach Learn Biomed Imaging.* 2024;2(2):888–915.
18. Otoo J, Nasiru S, Angbing ID. Shifted Hexpo activation function: An improved vanishing gradient mitigation activation function for disease classification. *Mach Learn Appl.* 2025;20(100651):100651.
19. Nakatumba-Nabende J, Murindanyi S. Deep learning models for enhanced in-field maize leaf disease diagnosis. *Mach Learn Appl.* 2025;20(100673):100673.
20. Lin Y-J, Chen C-A, Mao Y-C, Liang C-H, Chen T-Y, Li K-C, et al. Auxiliary evaluation of

marginal ridge discrepancy in periodontal disease using deep learning on periapical radiographs. Mach Learn Appl. 2025;(100727):100727.