



SvedbergOpen  
DISSEMINATION OF KNOWLEDGE

## International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

# Artificial Intelligence-Driven Computational Models And Biomathematical Algorithms For Predictive Healthcare Analytics, Disease Diagnosis, And Smart Medical Decision Support Systems

Dr.J. Refonaa<sup>1</sup>, Dr. Dinesh Suryakant Wankhede<sup>2</sup>, Dr. Sandeep R Gaigowal<sup>3</sup>, Dr. Sachin P. Nagmote<sup>4</sup>, Dr. Anjul Rai<sup>5</sup>

<sup>1</sup>Assistant Professor, Department of Civil Computer Science and Engineering, and Specialization in AI, Sathyabama Institute of science & Technology, Chennai, E-mail: [refonaa@gmail.com](mailto:refonaa@gmail.com)

<sup>2</sup>Associate Professor, Department of Electrical Engineering, St. Vincent Pallotti College of Engineering & Technology, Nagpur, ORCID: 0000-0002-6713-6020, E-mail: [dwankhede@stvincentngp.edu.in](mailto:dwankhede@stvincentngp.edu.in)

<sup>3</sup>Assistant Professor, Department of Electrical Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, E-mail: [sandeep\\_rg5@rediffmail.com](mailto:sandeep_rg5@rediffmail.com)

<sup>4</sup>Assistant Professor, Department of Electrical Engineering, St. Vincent Pallotti College of Engineering and Technology, Nagpur ORCID ID: 0009-0005-4187-5959, E-mail: [snagmote@stvincentngp.edu.in](mailto:snagmote@stvincentngp.edu.in)

<sup>5</sup>Assistant Professor, Department of Mechanical Engineering, Bhilai Institute of Technology (Chhattisgarh Swami Vivekanand Technical University (CSVTU), Durg, Chhattisgarh), Durg, Bhilai, ORCID id: 0009-0005-4187-5959, E-mail: [anjul.raibitdurg.ac.in](mailto:anjul.raibitdurg.ac.in)

### ABSTRACT

Cardiovascular disease remains a major public-health burden, requiring scalable and interpretable tools for early risk prediction and smart medical decision support. This study created an integrated framework based on artificial intelligence and biomathematical models to predict the risk of self-reported coronary heart disease (CHD) / myocardial infarction (MI) using the Behavioral Risk Factor Surveillance System (BRFSS) dataset 2024. Following pre-processing and outcome filtering, 452,464 respondents were included, and 9.36% of them had CHD/MI. Stratified train-test splitting, survey weighting, threshold optimization, probability calibration, explainability analysis and assessment of subgroups were used to train and evaluate four supervised models: weighted logistic regression, XGBoost, LightGBM, and CatBoost. CatBoost performed best in the held-out test set in terms of discrimination (AUROC = 0.8392; AUPRC = 0.3506). The CatBoost Brier score was reduced after Platt calibration from 0.1650 to 0.0712 and the final calibrated threshold was 0.08 with a sensitivity of 0.8250, a specificity of 0.6935, a negative predictive value of 0.9746, and a balanced accuracy of 0.7592. The Biomathematical Cardiovascular Risk Index was an interpretable comparator, with an AUROC of 0.7971. Clinically meaningful predictors were identified by SHAP analysis such as age, general health, diabetes, smoking, kidney disease and BMI. In summary, the calibrated CatBoost-BCRI is suitable for cardiovascular risk stratification and preventive decision support with interpretability, but needs further validation prior to clinical use.

**Keywords:** Artificial intelligence; Cardiovascular risk prediction; CatBoost; Biomathematical risk index; Smart medical decision support

## Introduction

Cardiovascular disease is one of the largest public health problems and is a major cause of mortality, morbidity, costs of healthcare and long term disease burden around the world. Epidemiological and biomathematical models have traditionally been used for risk estimation and they use a combination of demographic, behavior and clinical factors to estimate cardiovascular risk. In addition to highlighting the continued utility of mathematical risk models for 21 regions of the world, the revised World Health Organization cardiovascular disease risk charts also revealed that approaches need to be flexible to cater for the variability of risk across regions and population heterogeneity (Kaptoge et al., 2019). Heart disease and stroke remain a significant burden in the United States and around the world, and there is a need for scalable systems that can facilitate early detection, prevention, and risk-stratified health care planning (Martin et al., 2025).

AI and machine learning are also gaining more significance in predictive healthcare analytics, in addition to the traditional cardiovascular risk modeling. Such techniques are able to detect high-dimensional, non-linear associations among these multidimensional data, such as demographics, lifestyle indicators, access to healthcare, history of chronic diseases, self-reported health status, and missingness. Clinical expertise and artificial intelligence working together has been called 'high performance medicine' in which 'computational systems augment the clinician's work by improving prediction, diagnosis, workflow efficiency and personalized decision making' (Topol, 2019). Machine learning in medicine, in particular, can be helpful if there are predictor interactions in the data that follow a nonlinear pattern and a large number of predictor variables, which do not necessarily follow a linear pattern and a diverse range of patient types or patient-specific data that cannot be

fully described using a linear model (Rajkomar et al., 2019).

The adoption of AI in clinical and public health settings, though, requires some translation difficulties to overcome and make this technology a reliable method for decision-making. Predictive models can have very good discrimination, but be poorly calibrated, lack good explanations, and fail to perform well across subgroups. In healthcare, the key to a successful model is being actionable, interpretable, and reproducible, otherwise high numbers could be of little use for clinical value. The main challenges are dataset bias, sub-optimal external validation, insufficient calibration of the datasets, lack of transparency, and weak workflow integration and low clinician trust in the datasets (Kelly et al., 2019). In the context of cardiovascular risk prediction, these problems are of special significance: A false negative result can cause postponement of preventive treatment, while too many false positive results can result in unnecessary clinical burden and anxiety on the part of the patient.

To achieve responsible machine learning in healthcare, safeguards are necessary to reduce the impact of the models and to assess them by considering more than just overall accuracy. Sensitivity, specificity, negative predictive value, calibration, interpretability and consistency across subgroups should be assessed for a model designed for preventive screening. In medical machine learning, the “do no harm” principle has recently gained momentum, highlighting the need to build and manage predictive systems in a fair, safe, transparent and deployment context-sensitive (Wiens et al., 2019) way. Thus, cardiovascular risk prediction based on artificial intelligence should not be based solely on the black-box classification performance. Rather, it should combine calibrated probabilities, clinically interpretable risk groups, predictors that are understandable and biomathematical reasoning.

The present study is positioned within this intersection of artificial intelligence, biomathematical risk modeling, predictive healthcare analytics, disease diagnosis support, and smart medical decision-support systems. The research title “Artificial Intelligence-Driven Computational Models and Biomathematical Algorithms for Predictive Healthcare Analytics, Disease Diagnosis, and Smart Medical Decision Support Systems” outlines a paradigm that considers the machine learning models as a part of a larger decision-support system. In this context, gradient-boosting algorithms can be used to fine-tune predictions, calibration techniques can transform the output of these algorithms into probabilities that can be understood by the decision maker, biomathematical risk indices can offer a clear comparison of the scores, and explainability can pinpoint predictors within the algorithm that lead to the risk classification. This holistic view is bolstered by more recent advances in AI and machine learning for predictive medicine, particularly due to the increasing complexity of healthcare systems, which demand more accurate, scalable, interpretable, and adaptable models (Sharma et al., 2024).

The problem addressed in this study is the need for a predictive healthcare analytics framework that can identify individuals at elevated cardiovascular risk using large-scale population health data while maintaining interpretability and decision-support relevance. Traditional risk charts and statistical models are valuable tools and can't be replaced by something else, but may not fully leverage the structure of high dimensional survey-based health data. On the other hand, sophisticated models with more powerful AI can deliver better predictions but may not be well calibrated and transparent if not properly assessed. There is a methodological void because even though the use of AI-powered computational modeling, biomathematical risk scoring, probability calibration, explainability, and subgroup assessment have been established as cornerstones of cardiovascular decision support research, they are not yet integrated in a coherent manner.

The scope of this study is limited to secondary analysis of a large population health dataset and focuses on self-reported coronary heart disease/myocardial infarction status. The models are for predictive risk stratification and research in decision support and are not for direct diagnosis in the clinic. The findings of the survey should be looked at as evidence of the predictive value of the disease or outcome, rather than as confirmation of clinical adjudication of disease. Before embarking on real-world clinical use, prospective validation and external testing with datasets would be needed.

This study is significant because it proposes a structured computational framework that combines AI-based prediction, biomathematical risk-index development, probability calibration, calibrated risk grouping, and explainable decision support. The objectives of this study are:

- To develop and compare artificial intelligence-driven computational models for predicting self-reported coronary heart disease/myocardial infarction risk using large-scale population health data.
- To construct and evaluate a transparent biomathematical cardiovascular risk index as an interpretable comparator to machine learning models.
- To design a calibrated risk-stratification and smart medical decision-support framework using optimized thresholds, probability calibration, explainability analysis, and subgroup performance assessment.

The primary biomathematical contribution of this study is the definition and assessment of a clear cardiovascular risk function alongside calibrated machine-learning models. The Biomathematical Cardiovascular Risk Index is a convex weighted combination of the normalized clinical, behavioral, and demographic risk factors, and each of the risk factors can contribute explicitly to the final risk score. This structure allows direct comparisons of an interpretable mathematical scoring algorithm with flexible AI-based classifiers in the same survey-weighted predictive structure.

## 2. Literature Review

AI and machine learning are increasingly playing a significant role in cardiovascular risk prediction, diagnostic support in cardiovascular disease and predictive healthcare analytics. Machine learning can handle heterogeneous data sets with demographic, behavioral, clinical, healthcare-access, and longitudinal health variables, which are not used in conventional risk equations based on statistically defined relationships. AI-based cardiovascular disease risk prediction models perform well in a systematic review by Cai et al. (2024) but further validation in independent data sets, transparency in model development, and clinical utility is still necessary to successfully incorporate these models into clinical practice.

Recently, systematic reviews and meta-analyses have validated the increasingly important role of machine learning in cardiovascular prediction. Liu et al. (2025) reported that machine learning models using EHR data can enhance the estimation of cardiovascular risk, but highlighted ongoing challenges with quality reporting, calibration and external validation. Likewise, machine learning models have been used in various cardiovascular diseases demonstrating good prediction with some variability depending on the quality of the data, definition of the outcome, choice of predictors and validation protocol (Krittawong et al., 2020). The results suggest that AI models need to be thoroughly assessed rather than only for overall accuracy.

Machine learning has been shown to be valuable in the real-world setting of large cohort studies for population level risk prediction. Alaa et al. (2019) used automated machine learning algorithms on 423,604 participants of the UK Biobank to demonstrate that data-driven models can help with the estimation of cardiovascular risk at scale. Another study by Ward et al. (2020) showed that machine learning can be applied to predict atherosclerotic cardiovascular disease in a multi-ethnic cohort and that this predictive performance needs to be assessed for different demographic groups. The following studies endorse the application of sub-group analysis in decision-making research for cardiovascular diseases.

Despite the emergence of machine learning, the traditional cardiovascular risk score is still significant. While machine learning is superior to traditional scores in certain settings, traditional scores have benefits in terms of transparency, simplicity and familiarity to the clinical environment (Dimopoulos et al., 2018). This helps in an integrated approach where advanced AI models fine-tune predictive accuracy and biomathematical risk indices offer interpretable comparative benchmarks. The transparent cardiovascular risk index can, in this context, complement the use of machine learning by connecting the prediction with clinically meaningful weighted components.

AI is also known to be a game-changer in the field of cardiology. It has been described by Johnson et al. (2018) as a potential tool for cardiovascular imaging, phenotyping, risk prediction, prognosis and decision making, but with challenges in validation, interpretability and real world use. More generally, in the field of biomedical machine learning, reproducible data processing, feature selection, model validation and reporting of model performance are also emphasized (Luo et al., 2016). All of these factors play a pivotal role in creating believable predictive healthcare analytics models.

There is further evidence for methodological rigor in light of current frameworks for reporting and assessing bias. The TRIPOD+AI statement is focused on the transparency reporting of regression and machine learning prediction models, such as intended uses, calibration, and evaluation processes (Collins et al., 2024). PROBAST is a structured instrument to assess risk of bias and applicability within the four domains of participants, predictors, outcomes and analysis (Wolff et al., 2019). The two frameworks illustrate the need to report AUROC in addition to calibration, threshold selection, explainability, and subgroup evaluation in healthcare AI studies. The most important aspect of clinical decision support is the calibration of a model: it can provide a good ranking of patients, but it can also give incorrect probability estimates. Calibration is a significant drawback of predictive analytics as pointed out by Van Calster et al. (2019), and sound probability estimates are needed for any predictions used to make decisions. Calibrated probabilities are required for a cardiovascular screening to be able to put the people into clinically relevant risk groups.

Explainability is also key to medical AI adoption. Sadeghi et al. (2024) highlighted the importance of trust, accountability, and clinical acceptance in the context of explainable AI. Attribution-based approaches can help reveal the mechanisms of complex models making predictions, as demonstrated by Jin et al. (2022). In this context, such studies back the application of SHAP for clinical interpretation and highlight clinically-relevant features, like the general health, diabetes mellitus, smoking, kidney disease, and BMI. The literature overall provides support for an integrated approach for smart cardiovascular decision support based on machine learning, biomathematical scoring, calibration, explainability, and subgroup analysis.

## 3. Methodology

### 3.1 Study Design and Data Source

This study was designed as a retrospective predictive-analytics approach to create and test computational models, biomathematical algorithms, decision-support strategies, and calibrations that utilize artificial intelligence to predict cardiovascular disease risk. Data from the Centers for Disease Control and Prevention (Centers for Disease Control and Prevention, 2025) for the 2024 Behavioral Risk Factor Surveillance System (BRFSS) survey data and documentation were used for analysis. Original XPT data was processed and

converted to parquet files, which are suitable for analytics. The analytical data set consisted of 452,464 cases after removing invalid myocardial infarction/coronary heart disease outcome data.

The primary prediction target was obtained from the BRFSS variable \_MICHHD: Ever told they had CHD or MI. Valid responses were entered as either 1 (CHD/MI-positive) or 0 (no reported CHD/MI). This led to an unbalanced class distribution with 42338 positive cases and 410126 negative cases accounting for about 9.36% of positive prevalence.

### 3.2 Data Preprocessing and Variable Selection

Structured pre processing pipeline was developed prior to model development. Target leakage was avoided by excluding disease outcome variables that were directly associated with target from the predictor matrix, such as myocardial infarction, coronary heart disease, and stroke indicators. Predictors of candidates were chosen from demographic information, social determinants of health, access to health care, overall health status, behavioral risk factors, BMI, and chronic disease markers.

Where appropriate, response codes that were missing, refused, and uncertain for the BRFSS were recoded as missing. All continuous health-status variables, such as physical health days and mental health days, were recoded as 0 for the BRFSS value of "none." The BRFSS-derived \_BMI5 variable was used to calculate BMI and retained as a continuous predictor. To ensure robustness, selected smoking, alcohol, influenza, pneumococcal and prediabetes variables were removed from the primary model due to being high-missingness and/or unstable variables. For variables with meaningful amounts of missingness, missingness indicators were generated, and missingness was used as a potentially informative signal in the model.

After preprocessing, there were 66 predictors in the final set of modeling features: 44 primary predictors and 22 indicator variables for missing data. The continuous variables were median imputed while the categorical variables were imputed with the most frequent category. Categorical variables were one-hot encoded for logistic regression. Categorical variables are ordinal encoded or encoded natively in gradient boosting models.

### 3.3 Train-Test Split and Survey Weighting

Stratified sampling was used to ensure the original CHD/MI class distribution across the data set in order to split the data set into training and testing portions. There were 361,971 respondents in the training set and 90,493 respondents in the held-out test set. Both partitions maintained a target prevalence of ~9.36%.

Survey weights for the BRFSS survey were not dropped in the analysis because they were necessary for model estimation and evaluation, which are population-based. Primary sampling weight \_LLCPWT was standardized to its mean for model fitting to maintain the relative weights of the samples and to increase the numerical stability. Where applicable, in addition to the unweighted measures, survey-weighted measures such as weighted AUROC, weighted AUPRC, weighted accuracy, and weighted Brier score were reported.

### 3.4 Predictive Model Development

Four supervised binary classification models were trained and evaluated: weighted logistic regression, XGBoost, LightGBM and CatBoost. The statistical baseline used to interpret the results of the analysis was weighted logistic regression. The model employed L2 regularization, class-balanced weighting and normalized sample weights.

High performing gradient boosting models (XGBoost and LightGBM) were chosen for tabular healthcare data. Both models used class imbalanced processing and sample weighting. CatBoost was chosen due to the fact that there are many categorical predictors in the dataset, and CatBoost performs well with tabular data that consists mainly of categorical predictors. The CatBoost model is trained with weights based on the class, survey weights and an AUC-based evaluation procedure.

The model performances were evaluated based on AUROC, AUPRC, accuracy, balanced accuracy, precision, recall/sensitivity, specificity, F1-score, negative predictive value, Brier score and count of confusion matrix. Due to an imbalance in the results, along with AUROC, AUPRC, sensitivity and specificity were also highlighted.

### 3.5 Threshold Optimization and Probability Calibration

Given that the default value of 0.50 for the probability threshold of the default model is arbitrary for clinical screening, threshold optimization was done for all models. Both best F1 score and best balanced accuracy, the maximum Youden index, sensitivity  $\geq 0.80$  and maximum specificity, and sensitivity  $\geq 0.85$  and maximum specificity were used to evaluate thresholds. CatBoost's overall discrimination was the best and was chosen as the main AI model.

The probabilities of Raw CatBoost were subsequently calibrated using Platt scaling. The calibration procedure divided the prediction file held out into calibration and evaluation subsets. The raw CatBoost probabilities were first logit-transformed and then a logistic calibration model was fitted. Calibration performance was evaluated through calibration curves, Brier scores, weighted Brier scores and comparisons of the predicted risk versus the observed risk at the bin level. The calibrated CatBoost model was applied for risk-stratified decision support and for the interpretation of probabilities.

### 3.6 Biomathematical Cardiovascular Risk Index

In order to facilitate the biomathematical part of the study, an interpretable Biomathematical Cardiovascular Risk Index was created as a comparator to the AI models. The BCRI was computed as a composite of clinically interpretable normalized components, which included age, diabetes status, general health, BMI, smoking, physical health, physical inactivity, kidney disease, COPD and mental health. Scores for each of the components

was scaled from 0 to 1, with higher scores representing higher contribution to cardiovascular risk. The overall BCRI score was calculated as follows:

$$\text{BCRI} = 0.25(\text{age score}) + 0.15(\text{diabetes score}) + 0.15(\text{general health score}) + 0.10(\text{BMI score}) + 0.10(\text{smoking score}) + 0.07(\text{physical health score}) + 0.06(\text{physical inactivity score}) + 0.05(\text{kidney disease score}) + 0.04(\text{COPD score}) + 0.03(\text{mental health score}).$$

The weights were added together to give a total of 1.00 and assigned according to clinical plausibility, cardiovascular relevance, and ease of interpretation. The age, diabetes status and general health were given the heaviest weights due to their well established relationship with cardiovascular disease burden. BMI and smoking were given intermediate weights, as they have previously been shown to be associated with cardiovascular risk and relevant to assessment of preventive health. In addition, clinically relevant elements to consider for risk stratification were added (physical health, physical inactivity, kidney disease, COPD, and mental health). AUROC, AUPRC, weighted AUROC, weighted AUPRC, and threshold-based classification metrics were used as a measure of the evaluation of the BCRI. This index was used as an interpretable biomathematical comparator to the AI models, and was used to give an interpretable reference to assess if machine learning models outperform clinically weighted risk-scoring systems.

### 3.7 Explainability and Subgroup Analysis

CatBoost feature importance and SHAP values were used to assess model explainability. To allow identifying the most influential predictors, mean absolute SHAP values were calculated on a random sample of the test-set. A few high-risk and low-risk examples were also used to produce local SHAP explanations.

Subgroup performance analysis was performed by sex, age group, race, education, income, general health, diabetes status and BMI category. Subgroup metrics consisted of AUROC, AUPRC, sensitivity, specificity, precision, F1 Score, and numbers of confusion matrices. The analysis was used to determine model consistency and to identify threshold dependent variability among demographic, socioeconomic, and clinical groups.

## 4. Results

### 4.1 Analytical Dataset and Outcome Distribution

The final analytical sample for this analysis consisted of 452,464 respondents who had valid information for the coronary heart disease/myocardial infarction outcome from the BRFSS 2024 dataset after the data was preprocessed. The binary outcome variable of interest was created from the variable `_MICH` and it coded self-reported CHD/MI. Of those responding to the survey, 42,338 were CHD/MI-positive and 410,126 were CHD/MI-negative. This resulted in a relatively high positive-class prevalence of around 9.36%, meaning this was a highly imbalanced prediction problem. Due to this problem, the AUROC, AUPRC, sensitivity, specificity, negative predictive value, F1 score, Brier score, and the indicators of the confusion matrix were considered more significant than accuracy when evaluating models.

The final modeling data set consisted of 44 primary predictors and 22 variables used to indicate missingness in the data, giving 66 modeling features. These predictors included demographic, socioeconomic, health care access, general health status, behavioral risk factor, body-mass-index related factors, and chronic disease history. The data was split into a training set (361,971 respondents) and a testing set (90,493 respondents) using stratification. Both partitions had approximately equal CHD/MI positive prevalence of ~9.36%.

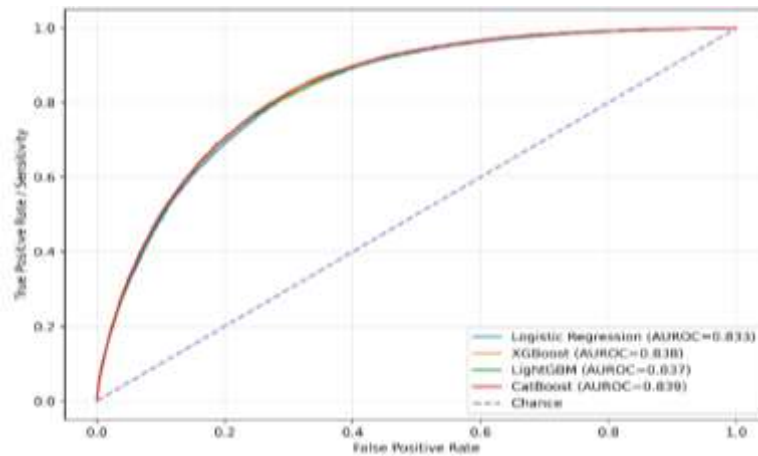
### 4.2 Comparative Discrimination of Predictive Models

Four predictive models were developed and compared: weighted logistic regression, XGBoost, LightGBM, and CatBoost. The held-out test-set discrimination results are presented in Table 1. CatBoost achieved the strongest overall discrimination, with an AUROC of 0.8392 and AUPRC of 0.3506. XGBoost followed closely, with AUROC of 0.8382 and AUPRC of 0.3499. LightGBM achieved AUROC of 0.8367 and AUPRC of 0.3476, while weighted logistic regression produced AUROC of 0.8332 and AUPRC of 0.3454. Although the differences among the gradient-boosting models were modest, CatBoost consistently achieved the highest discrimination values and was selected as the primary AI model.

**Table 1. Model discrimination comparison on the held-out test set**

| Model               | Test AUROC | Test AUPRC |
|---------------------|------------|------------|
| Logistic Regression | 0.8332     | 0.3454     |
| XGBoost             | 0.8382     | 0.3499     |
| LightGBM            | 0.8367     | 0.3476     |
| CatBoost            | 0.8392     | 0.3506     |

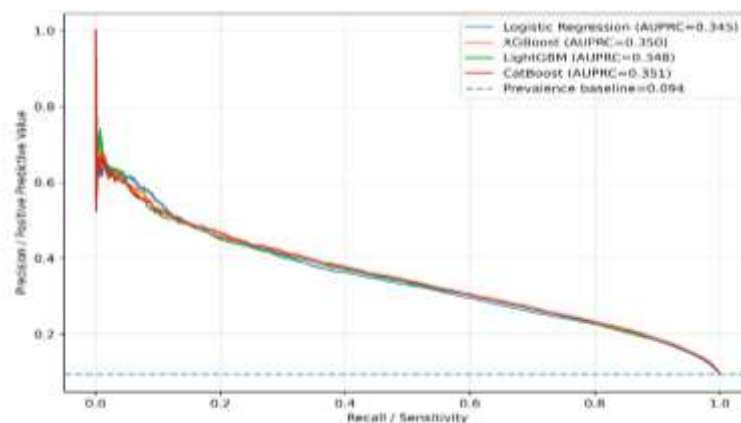
The ROC curves in Figure 1 show that all evaluated models performed clearly above chance level. CatBoost demonstrated the strongest curve profile overall, confirming its superior ranking ability for distinguishing CHD/MI-positive from CHD/MI-negative respondents.



**Figure 1. Receiver operating characteristic curves for all predictive models**

Receiver operating characteristic curves comparing weighted logistic regression, XGBoost, LightGBM, and CatBoost for CHD/MI risk prediction. CatBoost achieved the highest held-out test AUROC among the evaluated models.

The precision-recall curves in Figure 2 provide additional evidence under class imbalance. Since the CHD/MI-positive class represented less than 10% of respondents, AUPRC was particularly important. CatBoost achieved the highest AUPRC, indicating the strongest ability to prioritize CHD/MI-positive cases among predicted high-risk respondents.



**Figure 2. Precision-recall curves for all predictive models**

The dashed horizontal line represents the positive-class prevalence baseline. CatBoost achieved the highest AUPRC, indicating the strongest discrimination for the minority CHD/MI-positive class.

#### 4.3 Threshold Optimization for Decision Support

Threshold optimization was conducted because the default probability threshold of 0.50 is not necessarily appropriate for preventive healthcare screening. The selected decision-support criterion required sensitivity of at least 0.80 while maximizing specificity. The optimized threshold results are summarized in Table 2. Under this criterion, CatBoost achieved the best overall decision-support balance at a raw probability threshold of 0.48. At this threshold, CatBoost achieved sensitivity of 0.8043, specificity of 0.7212, precision of 0.2295, NPV of 0.9728, F1-score of 0.3571, and Youden index of 0.5255.

**Table 2. Decision-support threshold comparison using sensitivity  $\geq 0.80$  with best specificity**

| Model               | Threshold | AUROC  | AUPRC  | Accuracy | Sensitivity | Specificity | Precision | NPV    | F1     |
|---------------------|-----------|--------|--------|----------|-------------|-------------|-----------|--------|--------|
| Logistic Regression | 0.55      | 0.8332 | 0.3454 | 0.7214   | 0.8072      | 0.7126      | 0.2248    | 0.9728 | 0.3516 |
| XGBoost             | 0.48      | 0.8382 | 0.3499 | 0.7258   | 0.8039      | 0.7177      | 0.2272    | 0.9726 | 0.3543 |
| LightGBM            | 0.47      | 0.8367 | 0.3476 | 0.7265   | 0.8011      | 0.7188      | 0.2273    | 0.9722 | 0.3541 |

|          |      |        |        |        |        |        |        |        |        |
|----------|------|--------|--------|--------|--------|--------|--------|--------|--------|
| CatBoost | 0.48 | 0.8392 | 0.3506 | 0.7290 | 0.8043 | 0.7212 | 0.2295 | 0.9728 | 0.3571 |
|----------|------|--------|--------|--------|--------|--------|--------|--------|--------|

The CatBoost confusion-matrix counts at the optimized raw threshold included 59,154 true negatives, 22,871 false positives, 1,657 false negatives, and 6,811 true positives. These findings indicate that CatBoost was suitable for screening-oriented classification. However, raw probabilities required calibration before being interpreted as absolute disease-risk estimates.

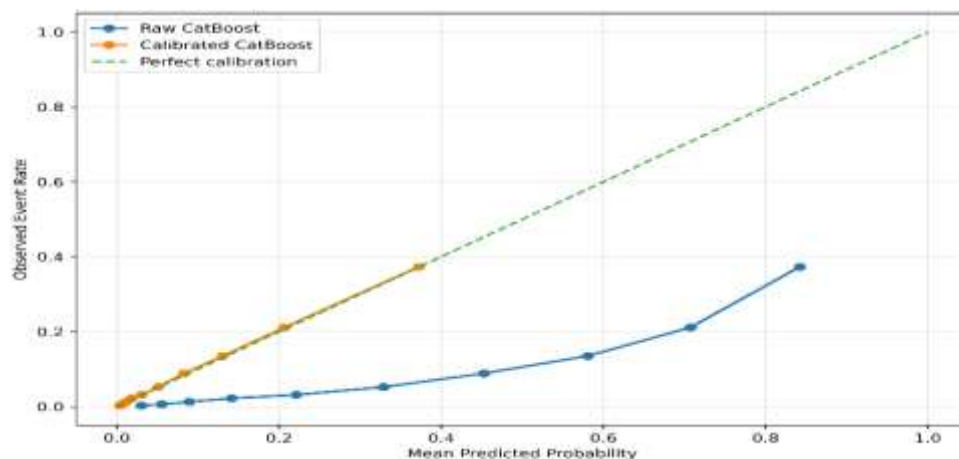
#### 4.4 Probability Calibration and Final Decision-Support Model

Calibration analysis showed that raw CatBoost probabilities substantially overestimated observed CHD/MI risk. Platt scaling was therefore applied to improve probability calibration. Calibration preserved AUROC and AUPRC because it did not change the ranking of individuals, but it markedly improved probabilistic accuracy. As shown in Table 3, the Brier score improved from 0.1650 to 0.0712, and the weighted Brier score improved from 0.1228 to 0.0536 after calibration.

**Table 3. Raw versus calibrated CatBoost probability metrics**

| Probability Type    | AUROC  | AUPRC  | Brier Score | Weighted Brier Score |
|---------------------|--------|--------|-------------|----------------------|
| Raw CatBoost        | 0.8353 | 0.3491 | 0.1650      | 0.1228               |
| Calibrated CatBoost | 0.8353 | 0.3491 | 0.0712      | 0.0536               |

The calibration curves in Figure 3 demonstrate the improvement after Platt scaling. Raw CatBoost predictions deviated strongly from the ideal calibration line, whereas calibrated predictions closely matched observed event rates across probability bins. This improvement supported the use of calibrated probabilities for risk communication and decision-support interpretation.



**Figure 3. Calibration curve for raw and calibrated CatBoost probabilities**

Calibration curve comparing raw CatBoost probabilities with Platt-scaled calibrated probabilities. Raw probabilities overestimated observed CHD/MI risk, whereas calibrated probabilities closely followed the ideal calibration line.

Using calibrated probabilities, the final decision-support threshold was set at 0.08. At this threshold, the calibrated CatBoost model achieved accuracy of 0.7058, weighted accuracy of 0.7865, balanced accuracy of 0.7592, precision of 0.2174, sensitivity of 0.8250, specificity of 0.6935, NPV of 0.9746, and F1-score of 0.3442. These results support its use as a preventive risk-stratification model rather than a definitive diagnostic instrument.

**Table 4. Final calibrated CatBoost decision-support performance at threshold 0.08**

| Model               | Threshold | Accuracy | Weighted Accuracy | Balanced Accuracy | Precision | Sensitivity | Specificity | NPV    | F1     |
|---------------------|-----------|----------|-------------------|-------------------|-----------|-------------|-------------|--------|--------|
| Calibrated CatBoost | 0.08      | 0.7058   | 0.7865            | 0.7592            | 0.2174    | 0.8250      | 0.6935      | 0.9746 | 0.3442 |

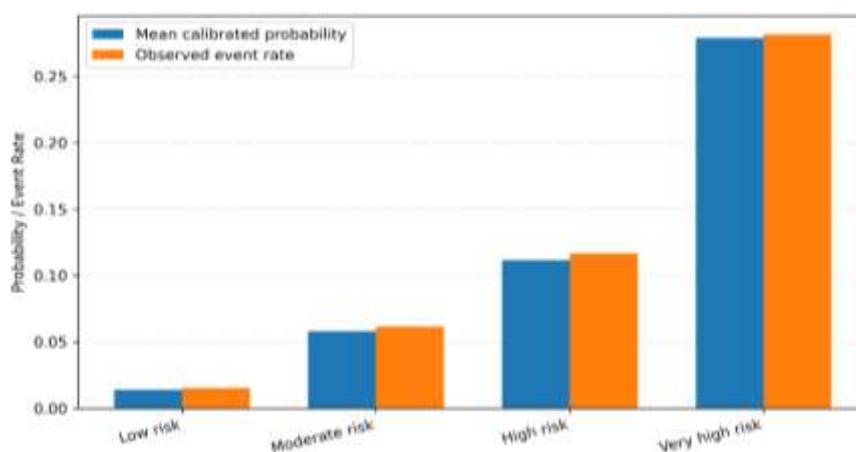
#### 4.5 Calibrated Risk Stratification

Calibrated CatBoost probabilities were used to define four risk groups: low risk, moderate risk, high risk, and very high risk. The risk-group summary is shown in Table 5. Observed CHD/MI event rates increased monotonically across groups, from 1.53% in the low-risk group to 28.15% in the very-high-risk group. This pattern supports the clinical interpretability of the calibrated risk-stratification system.

**Table 5. Calibrated CatBoost risk-group summary**

| Risk Group     | n     | Observed Cases | Mean Calibrated Probability | Observed Event Rate | Weighted Observed Event Rate |
|----------------|-------|----------------|-----------------------------|---------------------|------------------------------|
| Low risk       | 22801 | 348            | 0.0142                      | 0.0153              | 0.0135                       |
| Moderate risk  | 6382  | 393            | 0.0583                      | 0.0616              | 0.0605                       |
| High risk      | 6239  | 727            | 0.1116                      | 0.1165              | 0.1185                       |
| Very high risk | 9825  | 2766           | 0.2789                      | 0.2815              | 0.2777                       |

The predicted-versus-observed comparison in Figure 4 further supports calibration quality at the risk-group level. Mean calibrated probabilities closely matched observed event rates, demonstrating that the model provided clinically meaningful absolute risk estimates.

**Figure 4. Predicted versus observed risk across calibrated CatBoost risk groups**

Comparison of mean calibrated predicted probability and observed CHD/MI event rate across low, moderate, high, and very high risk groups. The close correspondence between predicted and observed risk supports the calibrated CatBoost model for risk-stratified decision support.

#### 4.6 Biomathematical Risk Index, Explainability, and Subgroup Robustness

Biomathematical Cardiovascular Risk Index was explored as an interpretable comparator. The BCRI results were obtained as AUROC of 0.7971 and AUPRC of 0.2825, and weighted AUROC of 0.8207 and weighted AUPRC of 0.2551. Its performance was below that of CatBoost, but the BCRI showed it to be a valuable risk-stratifier that uses clinically interpretable weighted components that are transparent.

Clinically relevant predictors identified by SHAP explainability were: age group, general health, personal doctor, employment status, sex, arthritis, diabetes, routine checkup, COPD, smoking history, kidney disease, physical health, age in years, race, and BMI. There was no significant difference in AUROC between sex groups, but more variation in AUROC between diabetes, general health, age and income groups. Calibrated CatBoost delivered the best discriminatory, calibration, sensitivity, negative predictive value and risk-stratification utility overall, while the BCRI helped to improve the biomathematical interpretability of the proposed decision-support framework.

## Discussion

The results of this study show that a process of merging artificial intelligence based and biomathematical methods can help to predict cardiovascular disease risks with large-scale population health data. CatBoost model had the highest held-out test discrimination (AUROC = 0.8392; AUPRC = 0.3506), marginally outperformed compared to XGBoost, LightGBM and weighted logistic regression. This finding indicates that gradient-boosting models are well suited to predicting cardiovascular risk by survey data, especially those that are able to process categorical-heavy tabular data. The improvement was not significant compared to logistic regression, but CatBoost demonstrated advantages in all the discrimination and decision-support metrics. In this way, the model contributed to the prediction and was compatible with the population health surveillance data structure.

The calibrated CatBoost model was the most useful for decision-support interpretation output. The average Brier score decreased from 0.1650 to 0.0712 and the weighted Brier score from 0.1228 to 0.0536 when using Platt scaling for the Raw CatBoost probabilities, which also predicted actual CHD/MI risk reasonably well. This is significant because a model can provide correct rankings for individuals, but incorrect probability

predictions. Calibrated probabilities are more meaningful than raw classification scores because they can be mapped to risk groups, a process of use for clinical/public-health decision support. The model was sensitive at 0.08 at 0.8250, specific at 0.6935, negative predictive value at 0.9746 and balanced accuracy at 0.7592. These values suggest that the model is more appropriate for preventive screening and/or risk stratification than definitive clinical diagnosis.

The findings are in line with recent reports in the literature on the potential gains in cardiovascular risk prediction using artificial intelligence models with large and heterogeneous health data sets. Cai et al. (2024) found that while AI-driven cardiovascular risk prediction models demonstrate potential, further robust validation, transparent reporting, and clinical interpretation are needed before their clinical use. In the current study, calibration, optimization of thresholds, explainability and subgroup analysis were included rather than just AUROC. Likewise, Liu et al. (2025) pointed out that large health data sets can be used to develop machine learning models for better cardiovascular risk estimation even though there are issues of inconsistent reporting, limited calibration assessment, and insufficient validation. The present framework addresses these concerns by examining the issue of discrimination, probability accuracy, risk stratification, and subgroup robustness.

The interpretability contribution of the CatBoost AI model versus the Biomathematical Cardiovascular Risk Index is important. The BCRI outperformed CatBoost with AUROC of 0.7971 and AUPRC of 0.2825, which shows reasonable, but less impressive discrimination power. Hence, transparent biomathematical scoring systems are able to reflect cardiovascular risk patterns that are relevant, whereas machine learning can model nonlinear relationships and complex feature interactions. Calibrated CatBoost is then more suitable as the ultimate predictive model to support decision making and the BCRI is more suitable as an interpretable comparator. This dual structure can be beneficial for medical AI systems as they should be predictive as well as clinically explainable.

The findings for explainability further corroborate the plausibility of the model. Some of the most important predictors were identified using SHAP analysis: age group, general health, access to a personal doctor, employment status, sex, arthritis, diabetes, routine checkup, COPD, smoking history, kidney disease, physical health, race, and BMI. These factors are similar to the known cardiovascular risk factors and social determinants of health. The application of the SHAP-based interpretation is aligned with the need for explainable healthcare AI, where explainability fosters trust, accountability, and clinical acceptance (Sadeghi et al., 2024). The AUROC was stable across sex groups; however, there was more variability across the age groups, income groups, general health groups, and diabetes groups. This implies that aggregate model performance is not adequate, and that the threshold-dependent behavior should be checked for each population subgroup.

There are a number of caveats that need to be noted. First, self-reported CHD/MI was used rather than clinically adjudicated cardiovascular events. Second, the lack of a baseline measurement prior to the intervention makes it difficult to interpret the results as causal effects. Third, the size of the data was large but the research was conducted by internal train-test validation, not external full validation using other data. Third, differences in subgroups can be due to disease prevalence, sample size differences, socioeconomic variations, or survey-response differences, as well as to algorithmic differences. These constraints are consistent with reporting concerns raised by prediction-models highlighted by Collins et al. (2024), such as transparency in reporting, validation, calibration and specification of intended use.

To confirm the calibration of the CatBoost model and BCRI, future studies need to externally validate these models on additional datasets, such as electronic health records and prospective cohorts of people undergoing cardiovascular screening. The next step should involve comparing the proposed framework to existing clinical cardiovascular risk scores, assessing the temporal validation of the framework, exploring different calibration methodologies, and exploring fairness-based thresholds. Future implementation research is also required to establish if there is evidence of enhanced preventive care, clinical prioritization and patient outcomes as a result of the use of calibrated risk stratification. In conclusion, the study validates the potential of AI models with calibrated algorithms and biomathematical models for predictive healthcare analytics and smart cardiovascular decision support.

## Conclusion

The purpose of this study was to develop and test an integrated, artificial intelligence-based and biomathematical framework for the prediction of cardiovascular disease (CVD) in a large-scale population health data set. CatBoost was the best of the three models tested in terms of discrimination on the held-out test set: better than weighted logistic regression, XGBoost, and LightGBM. The ultimate CatBoost decision-support model was clinically useful on the metrics of sensitivity, negative predictive value and balanced accuracy and was shown to be a tool for preventive risk stratification but not a diagnostic system. The Biomathematical Cardiovascular Risk Index also was predictive, and provided interpretability by assigning weight to clinical components of risk estimation. The results in this study suggest that smart medical decision-support systems could benefit from the combination of machine learning, probability calibration, biomathematical scoring, SHAP explainability, and subgroup evaluation. It has the potential to help healthcare professionals identify patients who might benefit from a more thorough cardiovascular risk assessment, preventive counseling,

lifestyle modification and clinical follow-up. But this model should be taken with pinches of prudence as it was a self-reported and internal validation. The externally testing of the calibrated CatBoost model and BCRI should be performed with new EHR datasets and with new screening cohorts in future studies. Future studies should also assess whether this framework can be compared to standard cardiovascular risk scores, optimize thresholds for each subgroup, explore different calibration methods and determine real-world clinical utility prior to implementation.

## Ethics Statement

Secondary data from the 2024 Behavioral Risk Factor Surveillance System (BRFSS) were used in this study and data were de-identified and publicly available. There is no personally identifiable information in the data set, nor was there direct contact with human participants in the analysis. Thus, no further institutional ethics clearance or informed consent was needed for this secondary data analysis.

## References

1. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & 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.
2. Cai, Y., Cai, Y. Q., Tang, L. Y., Wang, Y. H., Gong, M., Jing, T. C., ... & Zhang, G. W. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC medicine*, 22(1), 56.
3. Centers for Disease Control and Prevention. (2025). 2024 BRFSS survey data and documentation. U.S. Department of Health and Human Services. [https://www.cdc.gov/brfss/annual\\_data/annual\\_2024.html](https://www.cdc.gov/brfss/annual_data/annual_2024.html).
4. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385.
5. Dimopoulos, A. C., Nikolaidou, M., Caballero, F. F., Engchuan, W., Sanchez-Niubo, A., Arndt, H., ... & Panagiotakos, D. B. (2018). Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC medical research methodology*, 18(1), 179.
6. Jin, D., Sergeeva, E., Weng, W. H., Chauhan, G., & Szolovits, P. (2022). Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3), e1548.
7. Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., ... & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 71(23), 2668-2679.
8. Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M. T., Kavousi, M., Stevens, G., ... & Di Angelantonio, E. (2019). World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet global health*, 7(10), e1332-e1345.
9. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1), 195.
10. Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... & Tang, W. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, 10(1), 16057.
11. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). 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.
12. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ... & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research*, 18(12), e323.
13. Martin, S. S., Aday, A. W., Allen, N. B., Almarzooq, Z. I., Anderson, C. A., Arora, P., ... & American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Committee. (2025). 2025 heart disease and stroke statistics: a report of US and global data from the American Heart Association. *Circulation*, 151(8), e41-e660.
14. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
15. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., ... & Pardalos, P. M. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118, 109370.
16. Sharma, A., Lysenko, A., Jia, S., Boroevich, K. A., & Tsunoda, T. (2024). Advances in AI and machine learning for predictive medicine. *Journal of Human Genetics*, 69(10), 487-497.
17. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.

18. Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 17(1), 230.
19. Ward, A., Sarraju, A., Chung, S., Li, J., Harrington, R., Heidenreich, P., ... & Rodriguez, F. (2020). Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ digital medicine*, 3(1), 125.
20. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9), 1337-1340.
21. Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... & PROBAST Group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1), 51-58.