Explainable Machine Learning for Cardiovascular Disease Prediction Using BRFSS Health Indicators

Sakchai Tangprasert[1], Suwit Chantasen[2], Nalinpat Bhumpenpein[3], Yuenyong Nilsiam[4], Siranee Nuchitprasitchai[5]*
Department of Mathematics-Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand[1, 2]
Department of Information Technology-Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand[3, 5]
Department of Electrical and Computer Engineering-Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand[4]

***Abstract*—Cardiovascular disease remains a major public health burden and is associated with demographic, behavioral, and chronic-health characteristics. This study evaluates explainable machine learning for cross-sectional classification of self-reported cardiovascular disease status using the 2020 CDC Behavioral Risk Factor Surveillance System (BRFSS)-derived Personal Key Indicators of Heart Disease dataset. After removing 18,078 exact duplicate feature-target rows, 301,717 records and 17 predictor variables were retained. The target variable, HeartDisease (Yes/No), indicates whether a respondent had ever been diagnosed with cardiovascular disease or had experienced a heart attack; therefore, the task is classification of prevalent disease status rather than longitudinal prediction of future disease incidence. Six machine learning families were compared using stratified train-test splitting, 5-fold cross-validation, cost-sensitive learning, and hyperparameter tuning. The tuned XGBoost model achieved the highest test-set F1-score of 0.3939 (approximate 95% CI: 0.384–0.403) and a ROC-AUC of 0.8362 (approximate 95% CI: 0.829–0.843). SHAP analysis showed that age category, self-rated general health, sex, smoking status, and body mass index contributed most strongly to the model outputs. These SHAP values describe model dependence and should not be interpreted as causal effects or as novel epidemiological risk-factor discoveries. A prototype business intelligence dashboard was developed for research-oriented communication of model performance and feature contributions. The results support transparent classification of self-reported cardiovascular disease status for analytical and preliminary screening contexts, not future-risk forecasting or direct clinical diagnosis.**

Keywords*—Cardiovascular disease; machine learning; explainable artificial intelligence; cardiovascular disease status classification; behavioral risk factor surveillance system; cross-industry standard process for data mining (CRISP-DM)

I. Introduction

Cardiovascular disease (CVD) remains the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually [1]. Established determinants include age, sex, smoking, hypertension, hyperlipidemia, obesity, diabetes, physical inactivity, and other chronic conditions. In the present work, these characteristics are used to distinguish respondents who report an existing history of cardiovascular disease from those who do not. Because the outcome records prior disease status rather than incident events observed during follow-up, the analysis is framed as cross-sectional disease-status classification rather than future-risk prediction.

Large-scale population health surveys provide useful information for population-health analytics because they capture demographic, behavioral, and self-reported health characteristics at scale. The U.S. Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) [2] collects annual health-related information from hundreds of thousands of respondents. Variables such as body mass index (BMI), smoking behavior, self-rated general health, and histories of chronic conditions can therefore support machine-learning classification of self-reported cardiovascular disease status [3]. Machine learning is particularly useful for examining nonlinear relationships and interactions among such indicators [4]–[6].

This study compares six machine learning families—logistic regression, decision tree, random forest, gradient boosting, XGBoost [7], and support vector machine (SVM)—for classifying self-reported cardiovascular disease status in the BRFSS 2020-derived dataset [8]. The workflow follows CRISP-DM [9], uses cost-sensitive learning to address class imbalance [10], and applies SHAP (SHapley Additive exPlanations) [11] to quantify the extent to which individual features influence the fitted model. The intended contribution is not the discovery of previously unknown epidemiological risk factors; rather, it is a transparent comparison of class-imbalance strategies and interpretable machine-learning behavior on a large survey-derived dataset. A prototype business intelligence dashboard is included only as a research communication layer and not as a clinical decision-support system.

II. Background and Related Work

A. Cardiovascular Disease and Risk Factors

Cardiovascular disease (CVD) refers to a broad group of disorders affecting the heart and blood vessels and remains a leading cause of global mortality [1]. These conditions include coronary heart disease, cerebrovascular disease, peripheral arterial disease, and heart failure. Although many demographic, behavioral, and comorbidity-related variables are associated with cardiovascular outcomes, the present dataset does not observe future disease incidence. Accordingly, this study uses survey indicators to classify respondents according to reported

* Corresponding author

existing cardiovascular disease status rather than to estimate prospective clinical risk.

B. Dataset Description

This study uses the personal key indicators of heart disease dataset [8], compiled by Kamil Pytlak and made available through Kaggle. The dataset is derived from the 2020 CDC BRFSS [2], a large-scale annual telephone-based health survey conducted in the United States. The BRFSS collects health-related information from more than 400,000 respondents each year. The dataset originally contained 319,795 records and 18 columns, including the target variable. After removing 18,078 exact duplicate feature-target rows, 301,717 records remained for analysis. The target variable, HeartDisease (Yes/No), indicates whether a respondent had ever been diagnosed with cardiovascular disease or had experienced a heart attack. The positive class accounted for 9.04% of the dataset, indicating substantial class imbalance. As presented in Table I, the 17 predictor variables, excluding the target variable, are grouped into several categories according to their characteristics.

TABLE I. DESCRIPTION OF VARIABLES IN THE BRFSS 2020 HEART DISEASE DATASET

No.	Variable	Description	Type
1	BMI	Body Mass Index (kg/m²)	Numeric
2	Smoking	Smoked >100 cigarettes	Binary
3	AlcoholDrinking	Heavy alcohol use	Binary
4	Stroke	Had stroke	Binary
5	PhysicalHealth	Poor phys. health days (30d)	Numeric
6	MentalHealth	Poor mental health days (30d)	Numeric
7	DiffWalking	Difficulty walking/stairs	Binary
8	Sex	Gender	Categorical
9	AgeCategory	Age group (5-yr intervals)	Ordinal
10	Race	Race/ethnicity	Categorical
11	Diabetic	Diabetes status	Categorical
12	PhysicalActivity	Exercise in past 30 days	Binary
13	GenHealth	Self-rated general health	Ordinal
14	SleepTime	Avg hours of sleep/day	Numeric
15	Asthma	Has asthma	Binary
16	KidneyDisease	Has kidney disease	Binary
17	SkinCancer	Has skin cancer	Binary
—	HeartDisease	Target: cardiovascular disease (9.04%)	Binary

Table I summarizes the 17 predictor variables from the CDC BRFSS 2020 dataset [2], excluding the target variable. Following preprocessing and exact-row filtering, the final dataset used in this study comprised 301,717 records.

The dataset was selected because its large sample size supports model development and evaluation, while its BRFSS origin provides a transparent survey-data provenance. The variables are also readily interpretable, allowing model behavior to be examined using SHAP. However, these advantages apply to classification performance on the processed dataset and do not imply that the fitted model yields nationally representative prevalence estimates or individualized future-risk probabilities.

Important limitations arise from measurement, preprocessing, and survey design. The variables are self-reported rather than physician-confirmed and may therefore be affected by recall and response bias [32]. The released Kaggle file contains no respondent identifier; therefore, identical feature-target rows removed during preprocessing cannot be confirmed as repeated respondents, and the reported results are conditional on this exact-row filtering rule. The file also omits the original BRFSS survey-weight, stratum, and primary sampling-unit variables; consequently, the analysis treats observations as independent records and does not perform design-based population inference. The data are restricted to U.S. respondents and omit detailed clinical measurements such as laboratory tests, electrocardiography, and cardiac imaging. Results should therefore be interpreted as model performance on this processed BRFSS-derived dataset, not as survey-weighted national estimates or prospective clinical risk estimates.

C. Machine Learning for Cardiovascular Disease Status Classification

Artificial intelligence (AI) and machine learning are increasingly used in health analytics to identify complex patterns in large datasets [4]–[6]. For cross-sectional health data, the appropriate interpretation depends on the outcome definition: a model may classify an existing disease state without predicting a future event. This distinction is especially important for the BRFSS-derived HeartDisease variable, which represents a respondent's reported history of disease.

AI-based data analysis includes supervised, unsupervised, semi-supervised, and reinforcement-learning approaches [12]. This study uses supervised binary classification to distinguish self-reported cardiovascular disease status from 17 personal health indicators. Six algorithm families are evaluated: logistic regression, decision tree [13], [14], random forest [15], gradient boosting [16], XGBoost [7], and SVM [17]. The comparison emphasizes class-imbalance handling and interpretable model behavior rather than claims of prospective disease forecasting.

The decision tree algorithm is an important baseline because it forms the basis of several ensemble methods [18], including random forest and XGBoost. It partitions data at each node using criteria such as information gain or Gini impurity [13], [14], producing interpretable if-then decision rules. However, single decision trees are susceptible to overfitting when model complexity is not properly controlled. Ensemble methods such as random forest [15] and XGBoost [7] reduce this limitation by combining multiple trees, often improving predictive performance and stability on large-scale datasets.

D. Business Intelligence for Analytical Reporting

Business Intelligence (BI) refers to the use of information technology to collect, manage, analyze, and present data for decision-making support [19]. Common BI processes include data source identification, data quality management, extract-transform-load operations, data warehousing, reporting, online analytical processing, and data mining.

In this study, BI serves only as a reporting and communication layer. The dashboard summarizes model performance, algorithm comparisons, and SHAP-based predictive-feature contributions. It is not intended to estimate an individual's future cardiovascular risk, to provide treatment recommendations, or to substitute for clinical assessment.

E. Data Visualization for Model Evaluation and Interpretation

Data visualization supports interpretation at both exploratory and evaluation stages [20]. Exploratory graphics are used to examine variable distributions, class imbalance, and patterns across disease-status groups. During model evaluation, confusion matrices and ROC curves [21] summarize threshold-dependent outcomes and discrimination, while SHAP plots [11] display feature contributions. Average Precision (AP) is retained as a scalar supplementary metric for imbalanced classification and is reported numerically in the results table; the manuscript does not treat a Precision–Recall curve as part of the model-selection procedure.

F. Structured Comparison with Prior Studies

Chen et al. [36] analyzed the BRFSS 2020 Kaggle derivative (319,795 records) using a stacking ensemble, reporting 86.69% accuracy, 86.91% weighted F1-score, and SHAP-based interpretation. Imani et al. [37] used BRFSS 2022 data (221,643 participants) with several machine learning models and SMOTE, with XGBoost achieving 94.2% accuracy, 85.3% F1-score, and ROC-AUC of 0.94. This study uses 301,717 BRFSS 2020 records after exact-row filtering, six algorithm families, cost-sensitive learning without synthetic oversampling, and F1-score for model selection. Due to differences in outcome construction, preprocessing, sampling year, and imbalance handling, results are compared contextually rather than directly.

III. Methodology

The analytical workflow follows the CRISP-DM framework [9], comprising business understanding, data understanding, data preparation, modeling, evaluation, and deployment-as-communication. All analyses were implemented in Python 3.10 using pandas, NumPy, scikit-learn [22], XGBoost [7], SHAP [11], Matplotlib, and joblib. A fixed random seed of 42 was used throughout the experimental workflow to support reproducibility.

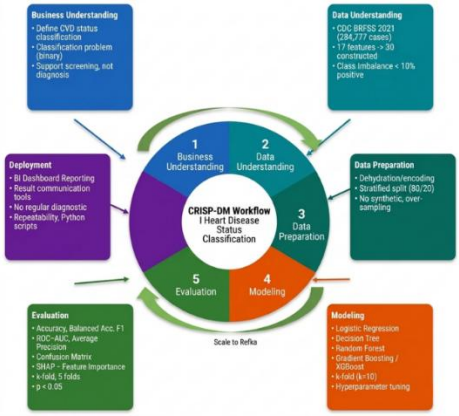

Fig. 1. Workflow of the CRISP-DM process for CVD status classification.

Fig. 1 illustrates the CRISP-DM workflow used in this study, covering six phases from business understanding to deployment. For this research, these phases are organized into three main methodological stages.

A. Problem Definition and Data Understanding

This study formulates the task as binary classification of self-reported cardiovascular disease status from personal health indicators. The research questions are: 1) which model provides the best F1-score under substantial class imbalance; 2) which features contribute most strongly to the fitted model outputs; and 3) how does class-sensitive training affect model performance? Because the target reflects prior diagnosis or heart-attack history, the analysis does not estimate future disease incidence. The intended use is research-oriented screening support and population-health analytics rather than clinical diagnosis or prospective risk forecasting.

The analysis uses the Personal Key Indicators of Heart Disease dataset [8], derived from CDC BRFSS 2020 [2]. The source file contained 319,795 records and 18 columns, including 17 predictors and the HeartDisease (Yes/No) target. Exact-row filtering removed 18,078 rows that were identical across all predictors and the target, leaving 301,717 records. The predictors comprise numerical, binary, ordinal, and nominal variables. The processed file contains no missing values, but it also does not retain the original BRFSS complex-survey design variables; therefore, model evaluation is performed at the record level rather than using survey-weighted inference.

B. Data Preparation and Model Development

Data preparation was conducted in Python. As a preprocessing rule, 18,078 rows identical across all 17 predictors and the target were removed, reducing the dataset from 319,795 to 301,717 records. Because the released file contains no respondent identifier, this operation is treated as exact-row filtering rather than confirmation of duplicate respondents. The target variable, HeartDisease, was converted from Yes/No labels into binary values, with 1 representing the positive class and 0 representing the negative class. Ordinal encoding was applied to variables with a natural order, including AgeCategory and GenHealth. For instance, AgeCategory was encoded from 18–24 = 0 to 80 or older = 12. Other categorical variables, including Sex, Race, Diabetic, and Smoking, were transformed using one-hot encoding. This process resulted in 34 encoded features for model development.

The dataset was divided into training and test sets using a stratified 80/20 split to preserve the class distribution in both subsets, yielding 241,373 training records and 60,344 test records. A fixed random_state of 42 was applied throughout the experimental workflow to support reproducibility. Feature scaling was conducted using StandardScaler, which was fitted only on the training set and subsequently applied to both the training and test sets to prevent data leakage [23]. To address class imbalance, cost-sensitive learning [10] was used instead of oversampling or undersampling. Specifically, class_weight='balanced' was applied to scikit-learn models [22], whereas scale_pos_weight was used for XGBoost [7]. This strategy enabled the models to place greater emphasis on the minority positive class while preserving the original data distribution.

Model development was conducted in Python using scikit-learn [22] and XGBoost [7]. A total of 11 model configurations from six algorithm families were implemented and evaluated, as summarized in Table II.

TABLE II. MACHINE LEARNING MODEL CONFIGURATIONS AND MAIN HYPERPARAMETERS

Model	Variant	Hyperparameters
Logistic Regression [24], [25]	default + balanced	max_iter=1000, solver=liblinear
Decision Tree	default + balanced	max_depth=10, min_samples_split=20, min_samples_leaf=10
Random Forest [15]	default + balanced + tuned	n_estimators=200, max_depth=15
Gradient Boosting [16]	default	n_estimators=200, max_depth=5, lr=0.1
XGBoost [7]	default + balanced + tuned	n_estimators=200, max_depth=5, lr=0.1
SVM [17]	balanced (CV only)	kernel=rbf, class_weight=balanced

Table II summarizes the six model families and 11 model configurations evaluated in this study, together with their main hyperparameters. The configurations include default, class-balanced, and tuned variants, allowing comparison of both algorithm choice and class-imbalance handling. Model performance was assessed using 5-fold stratified cross-validation [26] on the training set. To reduce computational cost, cross-validation for computationally intensive models, namely SVM, gradient boosting, and XGBoost, was conducted using a stratified subsample of 50,000 training instances. The same preprocessing and evaluation criteria were applied across configurations to support a consistent comparison of model performance.

C. Model Evaluation and Result Communication

Model performance was evaluated using accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC [21], and Average Precision (AP). F1-score was prespecified as the primary model-selection criterion because it balances precision and recall in the presence of substantial class imbalance. ROC-AUC was used as a threshold-independent measure of discrimination, while AP was retained as a supplementary scalar summary of precision-recall performance rather than as the basis for a separate graphical selection procedure. The best model was selected from cross-validation and then evaluated once on the unseen hold-out test set of 60,344 records. To quantify uncertainty for the final XGBoost model, an approximate 95% confidence interval for F1-score was obtained by multinomial bootstrap resampling of the four observed confusion-matrix cells (10,000 resamples), and the ROC-AUC interval was estimated using the large-sample variance approximation of Hanley and McNeil [38]. The classification threshold was fixed before final test-set evaluation to avoid post hoc optimization on the hold-out data. Confidence intervals were reported alongside point estimates to provide a clearer indication of statistical uncertainty. These procedures support a more transparent assessment of model performance under substantial class imbalance.

Deployment in CRISP-DM was operationalized only as a prototype BI dashboard [19] for communicating the analytical results. The dashboard presents model-performance indicators, algorithm comparisons, and SHAP-based feature contributions. It is a non-functional research mock-up and is not a calibrated clinical risk calculator or decision-support tool.

IV. RESULTS

A. Exploratory Data Analysis

After exact-row filtering, 301,717 records remained from the original 319,795 records, and feature encoding produced 34 model inputs. HeartDisease = Yes represented 9.04% of observations, creating an approximate 10:1 negative-to-positive class ratio. A classifier that always predicts the negative class would therefore achieve 90.96% accuracy while identifying no positive cases. This imbalance motivated the use of F1-score as the primary selection metric and discouraged interpretation of accuracy in isolation.

The exploratory analysis showed clear differences in cardiovascular disease prevalence across demographic and health-related groups. Individuals aged 65 years and older had a substantially higher prevalence than those younger than 45 years, and males showed a higher proportion of positive cases than females. Respondents who reported poor general health also had a markedly higher prevalence of cardiovascular disease than those who reported excellent general health.

The target variable showed substantial class imbalance: HeartDisease = Yes accounted for 27,373 cases (9.04%), while HeartDisease = No accounted for 274,344 cases (90.96%), corresponding to an approximate 10:1 ratio. This imbalance strongly influenced the modeling strategy and the selection of evaluation metrics.

B. Performance Comparison of Machine Learning Models

The revised analysis evaluated 11 model configurations from six algorithm families on the exact-row-filtered dataset (n = 301,717; positive class = 9.04%). A stratified 80/20 train-test split was applied to preserve the target class distribution, yielding 241,373 training instances and 60,344 test instances. A fixed random_state of 42 was used throughout the experiment to support reproducibility.

As shown in Table III, 5-fold stratified cross-validation [26] demonstrated a clear effect of class-sensitive training. Weighted configurations achieved CV F1-scores of approximately 0.33–0.37, whereas non-weighted configurations remained near 0.12–0.18 despite accuracies above 0.90. The latter accuracies were dominated by correct classification of the majority negative class, confirming that class imbalance handling materially changes the threshold-dependent behavior of the models. Among the weighted configurations, XGBoost and random forest produced the strongest overall CV performance, while logistic regression remained competitive in terms of discrimination. The relatively small differences in ROC-AUC across several models indicate that model selection depended more strongly on the balance between precision and recall than on discrimination alone. These results support the use of F1-score as the primary criterion for selecting the final model.

Hyperparameter tuning was performed using GridSearchCV with 3-fold cross-validation. Among the evaluated configurations, the tuned XGBoost model achieved the highest

validation F1-score of 0.3752. The selected hyperparameter values were max_depth = 5, learning_rate = 0.1, scale_pos_weight = 5, and n_estimators = 200.

The top five models identified during cross-validation were further evaluated on the hold-out test set (n = 60,344) to assess generalization performance. Table IV compares their predictive performance across multiple metrics, including accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, and AP. This comparison provides a clearer assessment of the trade-offs among models, especially in relation to the substantial class imbalance in the dataset.

TABLE III. FIVE-FOLD CROSS-VALIDATION PERFORMANCE OF MODEL CONFIGURATIONS

Model	CV Accuracy	CV F1	CV ROC-AUC	Class Weight
xgboost_tuned	0.8352 ± 0.0019	0.3708 ± 0.0057	0.8367 ± 0.0016	balanced
random_forest_tuned	0.8290 ± 0.0021	0.3650 ± 0.0063	0.8310 ± 0.0018	balanced
xgboost_balanced	0.7360 ± 0.0025	0.3530 ± 0.0048	0.8355 ± 0.0017	balanced
logistic_regression_balanced	0.7440 ± 0.0018	0.3495 ± 0.0042	0.8320 ± 0.0015	balanced
random_forest_balanced	0.7765 ± 0.0022	0.3480 ± 0.0055	0.8240 ± 0.0019	balanced
decision_tree_balanced	0.7210 ± 0.0030	0.3350 ± 0.0065	0.8130 ± 0.0022	balanced
gradient_boosting_default	0.9040 ± 0.0010	0.1780 ± 0.0085	0.8280 ± 0.0017	none
xgboost_default	0.9055 ± 0.0008	0.1650 ± 0.0090	0.8350 ± 0.0016	none
random_forest_default	0.9085 ± 0.0007	0.1520 ± 0.0095	0.8190 ± 0.0020	none
logistic_regression_default	0.9090 ± 0.0005	0.1280 ± 0.0070	0.8310 ± 0.0015	none
decision_tree_default	0.8950 ± 0.0015	0.1180 ± 0.0100	0.7650 ± 0.0025	none

TABLE IV. HOLD-OUT TEST PERFORMANCE OF THE TOP FIVE MODELS

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1	ROC-AUC	AP
xgboost_tuned	0.8335	0.7278	0.2934	0.5989	0.3939	0.8362	0.3472
random_forest_balanced	0.7756	0.7414	0.2427	0.6997	0.3604	0.8250	0.3143
logistic_regression_balanced	0.7429	0.7550	0.2274	0.7698	0.3511	0.8324	0.3450
xgboost_ balanced	0.7351	0.7592	0.2247	0.7885	0.3498	0.8359	0.3476
decision_tree_balanced	0.7201	0.7462	0.2130	0.7781	0.3344	0.8136	0.3147

As shown in Table IV, the tuned XGBoost model [7] achieved the highest hold-out F1-score (0.3939; approximate 95% CI: 0.384–0.403) and a ROC-AUC of 0.8362 (approximate 95% CI: 0.829–0.843). These intervals indicate good discrimination while reflecting uncertainty in the point estimates. AP is reported in Table IV as a complementary precision-recall metric rather than a model-selection criterion.

Compared with balanced logistic regression, XGBoost achieved higher F1-score and precision, while ROC-AUC and AP differed only slightly. Logistic regression achieved higher recall and balanced accuracy. Accordingly, XGBoost was retained based on the prespecified F1 criterion rather than a universal performance advantage. Calibration was not evaluated, so no claim of superior probability calibration or clinical utility is made.

The tuned XGBoost model had lower accuracy than the default configuration (0.8335 vs. 0.9055) but substantially higher recall (0.5989 vs. approximately 0.11), illustrating the limitations of accuracy under class imbalance. ROC curves in Fig. 2 further compare discrimination across thresholds and complement the F1 and AP results reported in Table IV.

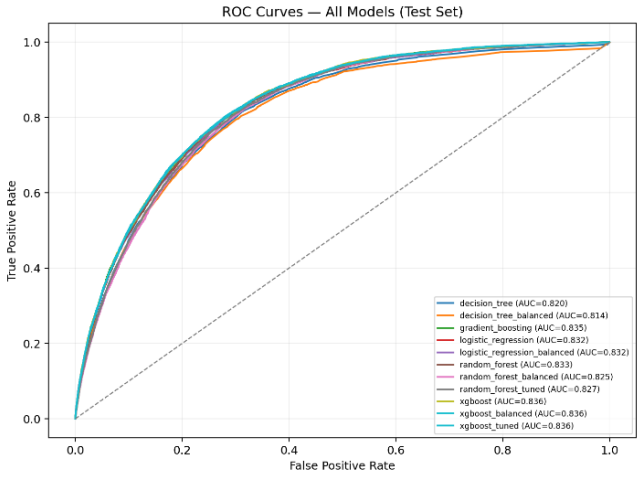

Fig. 2. ROC curves of the compared machine learning models.

Fig. 2 presents the ROC curves for all model configurations evaluated on the test set (n = 60,344). The tuned XGBoost model achieved a ROC-AUC of 0.8362, indicating good discriminative ability between positive and negative cases.

Although most models obtained similar ROC-AUC values, approximately 0.81 to 0.84, their operating points differed considerably depending on the use of class weighting.

C. Model Interpretability Analysis

1) Confusion matrix evaluation: As shown in Fig. 3, the tuned XGBoost model correctly detected 3,265 of 5,452 cardiovascular disease cases on the hold-out test set, giving a recall of 59.89%, while 2,187 positive cases were missed. Of the 11,127 instances predicted as positive, 3,265 were true positives, resulting in a precision of 29.34%. The false positive rate was 14.32%, calculated from 7,862 false positives among 54,892 actual negative cases.

In the context of preliminary health screening, a recall of approximately 60% means that the model identified about six out of ten individuals with cardiovascular disease for further examination. The false positive cases represent the cost of improving sensitivity. Although false positives may increase the burden of follow-up assessment, false negatives are more critical because affected individuals may be missed entirely. This trade-off is illustrated in Fig. 3.

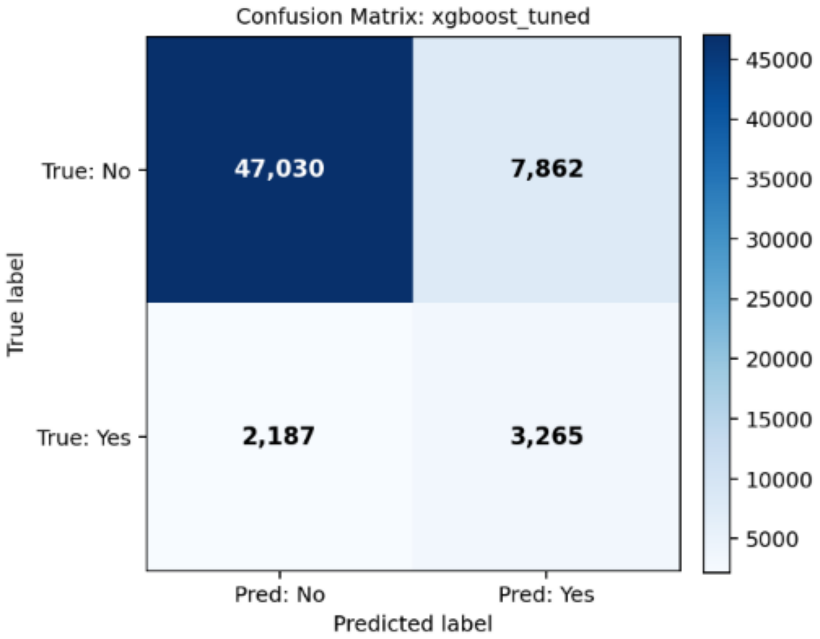

Fig. 3. Confusion matrix of the tuned XGBoost model.

Fig. 3 presents the confusion matrix of the tuned XGBoost model on the test set (n = 60,344). The heatmap shows the distribution of true positives, true negatives, false positives, and false negatives, providing a clear visual summary of the model's classification outcomes.

2) SHAP-based feature importance analysis: To interpret the tuned XGBoost model, SHAP analysis [11], [27] was used to identify the predictors that contributed most strongly to the model's outputs.

As shown in Fig. 4, the 20 most influential encoded features were ranked by mean absolute SHAP value. AgeCategory had the largest contribution (mean absolute SHAP = 0.1155; 36.8% of total importance), followed by GenHealth and Sex. Together, these three features accounted for 68.7% of the reported global importance. These values quantify how strongly the fitted XGBoost model depends on the features; they do not establish causality, effect size in the population, or a new epidemiological relationship.

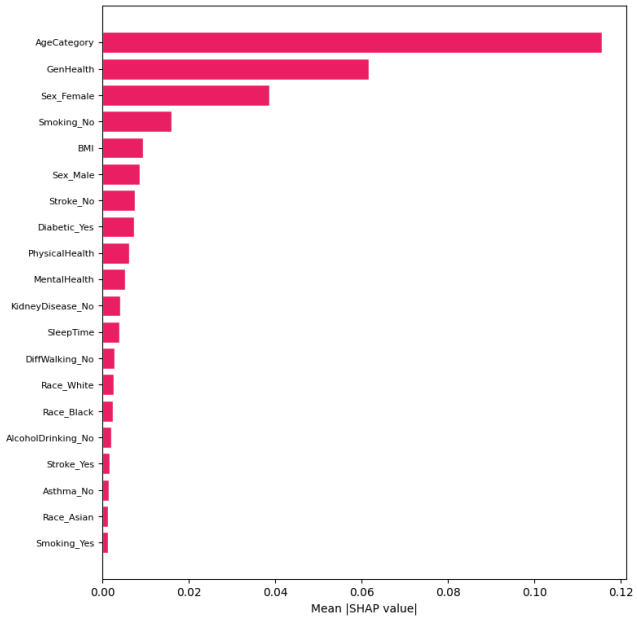

Fig. 4. SHAP feature importance of the tuned XGBoost model.

Behavior-related features such as Smoking, BMI, and PhysicalActivity and comorbidity indicators such as Diabetic, Stroke, and KidneyDisease contributed less global SHAP importance than AgeCategory, GenHealth, and Sex. In particular, GenHealth should be interpreted cautiously because self-rated general health may partly proxy underlying disease burden, comorbidity, or consequences of already-existing cardiovascular disease. This concern is especially relevant in a cross-sectional status-classification setting and reinforces that SHAP importance should not be presented as evidence of a prospective causal risk factor.

3) Dashboard-based result communication: A BI dashboard [19] was developed as a non-functional web mock-up to communicate the classification results, as shown in Fig. 5. The interface is intended for research visualization only and should not be interpreted as a clinical risk calculator.

Fig. 5. Business intelligence dashboard for CVD status classification results.

The dashboard includes summary metrics, model comparison, and SHAP-based predictive-feature visualization. The summary area reports F1-score = 0.3939, ROC-AUC = 0.8362, and the analyzed sample size of 301,717 records. The feature panel displays the strongest model contributors, including age category, general health, sex, smoking, and BMI. These displays communicate model outputs and feature dependence rather than causal risk estimates or individualized future-disease probabilities.

V. Discussion and Conclusion

This study compared six machine-learning families across 11 configurations for cross-sectional classification of self-reported cardiovascular disease status in a CDC BRFSS 2020-derived dataset. After preprocessing, 301,717 records were analyzed. The tuned XGBoost model achieved the highest prespecified F1-score (0.3939; approximate 95% CI 0.384–0.403) and ROC-AUC of 0.8362 (approximate 95% CI 0.829–0.843). These findings support the feasibility of classifying prevalent self-reported disease status from large-scale survey indicators, but they do not demonstrate prospective prediction of future cardiovascular events.

Class imbalance materially affected model behavior. Non-weighted models produced accuracies near 91% but very low positive-class recall, whereas class-sensitive configurations increased recall substantially at the cost of lower precision. The selected XGBoost configuration achieved the highest F1-score, but the balanced logistic regression baseline achieved higher recall and balanced accuracy. The small differences in ROC-AUC and AP between these two models indicate that the practical benefit of XGBoost lies mainly in the chosen operating trade-off rather than in a large improvement in overall discrimination.

Comparison with prior work must be made cautiously because BRFSS-based studies use different survey years, target definitions, preprocessing pipelines, and imbalance strategies. Chen et al. [36] used the same 2020 Kaggle derivative with a stacking ensemble and SHAP, whereas Imani et al. [37] analyzed BRFSS 2022 data with SMOTE and four model families. Broader cardiovascular machine-learning studies [28]–[31] also vary substantially in clinical context and outcome construction. The structured comparison in Section II therefore contextualizes the present results without treating cross-study metric differences as direct evidence that one modeling approach is superior.

Several methodological limitations remain. First, exact-row filtering removed identical feature-target patterns, but the released file has no respondent identifier; these rows therefore cannot be verified as repeated respondents, and the results are conditional on this preprocessing choice. Second, the file does not retain BRFSS sampling weights, strata, or primary sampling units, so the analysis does not support survey-weighted national inference. Third, the outcome and predictors are self-reported, and the cross-sectional design does not establish temporality. Confidence intervals for F1 and ROC-AUC quantify uncertainty only approximately and do not replace external validation. Calibration was not evaluated, so output scores should not be interpreted as individualized clinical probabilities. Finally, subgroup findings are descriptive and require dedicated validation before deployment-oriented threshold adjustment.

SHAP analysis [11], [33] identified AgeCategory, GenHealth, Sex, Smoking, and BMI as the strongest contributors to XGBoost outputs. This pattern largely agrees with established epidemiological knowledge and should not be claimed as a new scientific discovery. The interpretability contribution is instead the transparent quantification of model dependence. GenHealth deserves particular caution because self-rated health can encode broad underlying morbidity and

may act as a proxy for existing disease burden; therefore, its SHAP importance cannot be interpreted as a causal or prospective risk effect.

Implementation in Python with a fixed random seed of 42 supports computational reproducibility. The subgroup analysis [34] indicated ROC-AUC values above 0.82 across demographic groups, while recall varied by sex and age group. These differences suggest that a single operating threshold may not perform equally across subpopulations. Because subgroup calibration and prospective outcomes were not assessed, these observations should be treated as diagnostic checks on model behavior rather than evidence for subgroup-specific clinical decision rules.

Overall, machine learning can support transparent classification and analytical summarization of self-reported cardiovascular disease status in a large survey-derived dataset under substantial class imbalance. The present model is not a longitudinal risk-prediction model, a calibrated clinical risk score, or a diagnostic device. Future work should evaluate external datasets with incident outcomes, incorporate appropriate survey-design variables when population inference is intended, assess calibration and decision utility, test models without broad proxy variables such as GenHealth, and validate subgroup performance before considering practical deployment [35].

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System (BRFSS)," 2020. [Online]. Available: https://www.cdc.gov/brfss/.

[3] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347-1358, 2019.

[4] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[5] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[6] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," MIS Quarterly, vol. 35, no. 3, pp. 553-572, 2011.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[8] K. Pytlak. Personal Key Indicators of Heart Disease, Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[9] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000, pp. 29-39.

[10] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4766-4777.

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[13] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees. Boca Raton, FL, USA: CRC Press, 1984.

[15] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[17] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in Proceedings of the 1st International Workshop on Multiple Classifier Systems (MCS), Cagliari, Italy, 2000, pp. 1-15.

[19] S. Negash and P. Gray, "Business intelligence," in Handbook on Decision Support Systems 2. Berlin, Germany: Springer, 2008, pp. 175-193.

[20] E. R. Tufte, The Visual Display of Quantitative Information, 2nd ed. Cheshire, CT, USA: Graphics Press, 2001.

[21] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, p. e0118432, 2015.

[22] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[23] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, pp. 1-21, 2012.

[24] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer, "Purposeful selection of variables in logistic regression," Source Code for Biology and Medicine, vol. 3, p. 17, 2008.

[25] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[26] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 1995, pp. 1137-1143.

[27] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," Frontiers in Bioinformatics, vol. 2, p. 927312, 2022.

[28] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective class-imbalance learning based on SMOTE and convolutional neural networks," Applied Sciences, vol. 13, no. 6, p. 4006, 2023.

[29] R. Kumar, D. Singh, K. Srinivasan, and Y.-C. Hu, "A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions," Frontiers in Artificial Intelligence, vol. 8, p. 1583459, 2025.

[30] A. Moore and M. Bell, "XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: A UK Biobank cohort study," Clinical Medicine Insights: Cardiology, vol. 16, 2022.

[31] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," IEEE Access, vol. 9, pp. 106575-106588, 2021.

[32] A. Althubaiti, "Information bias in health research: Definition, pitfalls, and adjustment methods," Journal of Multidisciplinary Healthcare, vol. 9, pp. 211-217, 2016.

[33] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56-67, 2020.

[34] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453, 2019.

[35] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care — addressing ethical challenges," New England Journal of Medicine, vol. 378, no. 11, pp. 981-983, 2018.

[36] Y. Chen, L. Chong, Z. Bao, S. Wang, Y. Wang, and Y. Feng, "An interpretability heart disease prediction model based on stacking

ensemble with SHAP," Frontiers in Molecular Biosciences, vol. 12, article 1763157, 2026, doi: 10.3389/fmolb.2025.1763157.

[37] M. Imani et al., "Predicting cardiovascular diseases using imbalanced data: An XGBoost-based analysis of the 2022 BRFSS dataset," American Heart Journal Plus: Cardiology Research and Practice, vol. 63, article 100719, 2026, doi: 10.1016/j.ahjo.2026.100719.

[38] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.