# Cardiovascular Disease Detection Based on Lifestyle Factors

Dmitri Lezama
Dept. of Computer Science & Information Technology
The University of the West Indies
St. Augustine, Trinidad and Tobago
dmitri.lezama@my.uwi.edu

Lorenzo Gould-Davies
Dept. of Computer Science & Information Technology
The University of the West Indies
St. Augustine, Trinidad and Tobago
lorenzo.goulddavies@my.uwi.edu

Felicia Chan
Dept. of Computer Science & Information Technology
The University of the West Indies
St. Augustine, Trinidad and Tobago
felicia.chan@my.uwi.edu

Sergio Mathurin
Dept. of Computer Science & Information Technology
The University of the West Indies
St. Augustine, Trinidad and Tobago
sergio.mathurin@sta.uwi.edu

***Abstract*—This study addresses cardiovascular disease (CVD) risk prediction through machine learning application on the CDC's Behavioral Risk Factor Surveillance System 2023 dataset. We implemented a systematic feature selection methodology, reducing 350 variables to 18 clinically relevant predictors. Seven machine learning models were evaluated using metrics prioritizing both predictive accuracy and clinical utility. Gradient-boosted frameworks demonstrated superior performance, with LightGBM achieving an F1 score of 82.6% and ROC AUC of 87.5%. Age, high cholesterol, mobility impairments, and BMI were identified as principal predictors. A custom interpretability algorithm was developed to replace SHAP, maintaining feature attribution integrity while reducing computational overhead. The final model integrates the LightGBM classifier with a large language model for translating statistical outputs into actionable recommendations. Post-development validation on self-reported data demonstrated substantial agreement with manual annotations (Cohen's $\kappa = 0.92$). This research establishes a methodological framework applicable to other chronic condition risk assessments and contributes to more targeted preventive interventions in public health contexts.**

***Index Terms*—Cardiovascular Disease, Risk Prediction, Machine Learning**

## I. Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, accounting for approximately 18 million deaths annually and imposing substantial socioeconomic burdens through healthcare costs and lost productivity. Despite advances in medical science, the identification of individuals at risk for CVD continues to present significant challenges in clinical practice. This research addresses the critical need for accurate, interpretable predictive models that can effectively stratify cardiovascular risk using widely available health data.

This study leverages the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) 2023 dataset to develop and evaluate probabilistic classification models for CVD risk assessment. The BRFSS represents one of the most comprehensive population-based health surveys available, providing structured data suitable for machine learning applications while avoiding the privacy constraints that typically limit access to clinical datasets.

We apply a systematic feature selection process, reducing an initial set of 350 variables to 18 non-redundant predictors with established clinical relevance to cardiovascular health. Seven probabilistic machine learning models were tested, ranging from linear classifiers to gradient boosting models and multilayer perceptrons. These models were evaluated using metrics that prioritize both predictive accuracy and clinical utility, with particular focus on the F1 score and ROC AUC to minimize false positives and false negatives in cardiovascular risk assessment.

Additionally, this research addresses the ”black box” problem common in predictive healthcare analytics by incorporating a custom interpretability framework. This approach enables transparent feature attribution while maintaining computational efficiency, allowing for real-time explanation of model predictions. The integration of these interpretable predictions with a large language model for generating personalized risk insights represents a novel contribution to clinical decision support system design.

The findings presented demonstrate the enhanced performance of gradient-boosted frameworks for CVD prediction when applied to population-level survey data. The methodological framework established may serve as a template for developing similar risk prediction tools for other chronic conditions, potentially contributing to more efficient resource allocation and targeted preventive interventions in public health contexts.

## II. Background

Despite considerable medical advancements, cardiovascular disease (CVD) persists as the world's primary cause of mortality, claiming approximately 18 million lives annually. [1], [2] This imposes severe socioeconomic costs through healthcare burdens and lost productivity [3]. In 2022, reported CVD deaths reached 19.4 million, exceeding the combined mortality of cancer (9.7 million) and chronic lower respiratory diseases (2.6 million) [4]–[6]. These conditions affect the heart and blood vessels, with the most common types being ischemic heart disease, cerebrovascular disease (stroke), hypertensive heart disease, cardiomyopathy, and myocarditis [7]. Of these conditions, ischemic heart disease and stroke account for 85% of this epidemiological burden [8].

### *A. Risk Factors for Cardiovascular Disease*

Primary risk factors have been established through decades of epidemiological research, falling into three categories: behavioral, environmental, and metabolic risks. According to a statistical report [7], behavioral risks such as poor dietary habits, tobacco use, and excessive alcohol consumption are among the most significant contributors to CVD. Metabolic factors, namely, hypertension, high cholesterol, and obesity, further elevate CVD risk. These findings are consistent with studies presented in [9], [10]. Type 2 diabetes is particularly critical, as it can increase the likelihood of coronary heart disease and ischemic stroke by up to 4 times while also significantly raising mortality rates [11]. Growing evidence highlights several additional contributors to CVD risk. Sensory impairments, such as hearing and vision loss, are recognized as indirect but important indicators of CVD vulnerability due to their association with common CVD predictors like diabetes, smoking, hypertension, and obesity [12], [13]. Physical inactivity also plays a major role, with studies demonstrating increased hazard ratios across various CVD conditions [14]. Similarly, impairments in physical mobility have been strongly linked to increased CVD risk [15]. Preventative measures such as vaccination have been shown to reduce residual cardiovascular risk [16], highlighting the multifaceted nature of CVD prevention.

### *B. Data Availability for CVD Research*

The development of advanced predictive tools for CVD is informed by several available datasets, each presenting unique advantages and limitations. Among these, the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) is a significant resource, offering data from a large-scale, nationwide medical survey conducted by professional statisticians. This system provides structured and accessible health information suitable for cardiovascular research and machine learning development [17].

Other data sources, such as the IEEE dataset [18], Mendeley dataset [19], and various Kaggle datasets [20]–[22], also contribute to research efforts; however, they often exhibit characteristics like smaller sample sizes, fragmentation, or the use of synthetically augmented data, which can potentially lead to less representative findings and complicate the development of generalizable models. Even the BRFSS dataset has limitations; for instance, recent policy changes, such as those enacted during the Trump administration, have resulted in the withholding of certain features, thereby posing challenges to cross-year data collation.

### *C. Machine Learning Approaches for CVD Prediction*

Investigations into the efficacy of machine learning for heart disease prediction include comparative studies of various algorithmic approaches. The study presented in [23], for example, conducted an analysis evaluating five distinct machine learning algorithms: Random Forest, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbor Classifier, and a Decision Tree Classifier optimized with Grid search. Their research, which also involved the development of an ensemble model and utilized a heart disease dataset from Kaggle, reported that the Decision Tree Classifier achieved the highest F1-score at 93%. However, a potential limitation of this study is the relatively small dataset size of only 303 records, which may affect the generalizability and robustness of the reported model performance.

The challenge of class imbalance in heart disease datasets and its impact on diagnostic accuracy has also been a focus of research. Other papers, [24] addressed this by employing the Synthetic Minority Over-sampling Technique (SMOTE) on a 2022 BRFSS dataset obtained from Kaggle. Their study evaluated ten machine learning algorithms, incorporating hyperparameter tuning and 5-fold cross-validation, and assessed performance using metrics such as accuracy, precision, sensitivity, F1-score, and ROC-AUC. While XGBoost demonstrated the highest F1-score at 91.7%, Logistic Regression and Artificial Neural Networks also yielded comparable and satisfactory performance, achieving F1-scores of 91.5% and 91.4% respectively, highlighting the potential of these models in enhancing early heart disease detection. However, the study's reproducibility was suboptimal due to limited methodological detail.

## III. Methodology

### *A. Data Preprocessing*

As an initial step, variable names from the CDC BRFSS Calculated Variables 2023 code-book were standardized and mapped to more descriptive labels to enhance interpretability and readability. In addition, categorical variable were reencoded to formats better suited for machine learning workflows, such as converting ordinal and nominal variables to integer representations, thereby improving compatibility with downstream modeling algorithms.

A consolidated cardiovascular indicator variable (denoted as ”CVD”) was created by aggregating multiple cardiovascular disease diagnoses into a single binary outcome. This simplification was chosen to prioritize early detection of any heart-related condition, rather than distinguishing among specific subtypes.

Records containing missing values were removed, eliminating 110,227 entries and reducing the dataset size from 433,016 to 322,789 instances. This approach avoided model bias, and imputation showed no performance gain in empirical testing. Several variables in the dataset were pre-constructed from imputed components as part of the CDC's preprocessing, reducing opinionated feature merging during data preparation.

### *B. Feature Selection*

Feature selection reduces dimensionality by isolating predictors that strongly influence outcomes, thereby improving predictive accuracy and inference efficiency. A literature-informed filter reduced the initial 350 candidate features to 31 variables with well-documented correlations to cardiovascular disease. During this process, semantically related features were merged to eliminate redundancy: difficulty walking, dressing, bathing, and running errands were consolidated into mobility impairments; blind or vision difficulty and deaf or hard of hearing were combined into sensory impairments; and flu and pneumonia-vaccinated were aggregated into vaccination status.

TABLE I
FEATURES DERIVED FROM BRFSS 2023 DATASET

| FEATURE | ENCODING |
|---|---|
| SEX | Numeric [0=male, 1=female] |
| AGE | Categorical [1 to 13] |
| HEIGHT | Numeric [0.90 to 2.50] |
| WEIGHT | Numeric [20 to 300] |
| BMI | Numeric [0 to 1] |
| GENERAL HEALTH | Categorical [5 levels] |
| PHYSICAL ACTIVITY | Numeric [0 to 1] |
| ASTHMA | Numeric [0 to 1] |
| COPD | Numeric [0 to 1] |
| KIDNEY DISEASE | Numeric [0 to 1] |
| ARTHRITIS | Numeric [0 to 1] |
| DIABETES | Categorical [3 levels] |
| SMOKER STATUS | Numeric [0 to 1] |
| ALCOHOL CONSUMPTION | Numeric [0 to 1] |
| HIGH CHOLESTEROL | Numeric [0 to 1] |
| SENSORY IMPAIRMENTS | Numeric [0 to 1] |
| MOBILITY IMPAIRMENTS | Numeric [0 to 1] |
| VACCINATION STATUS | Numeric [0 to 1] |
| CVD | Numeric [0 to 1] |

To further refine the feature set, a coarse-to-fine search strategy was leveraged to eliminate potentially redundant or weakly informative features. In the coarse phase, features with low empirical variance (>0.05) or minimal uni-variate correlation (p>0.2) with cardiovascular diseases were eliminated. Specifically, a feature $f_i \in \mathcal{F}$ was retained only if it satisfied:

$$\mathrm{Var}(f_i) > \tau_v \quad \text{and} \quad |\rho(f_i, y)| > \tau_c,$$

where $\tau_v = 0.05$ is the variance threshold and $\tau_c = 0.2$ is the minimum Pearson correlation coefficient with the target $y \in \{0, 1\}$. This filtering step removed low-information and noise-prone variables.

Following this, the remaining feature pool $\mathcal{F}_{\text{coarse}}$ underwent a fine-grained evaluation. All subsets $\mathcal{S} \subseteq \mathcal{F}_{\text{coarse}}$ were assessed via model training to identify the combination that maximized predictive performance. The final feature set $\mathcal{F}^*$ was selected by solving:

$$\mathcal{F}^* = \arg \max_{\mathcal{S} \subseteq \mathcal{F}_{\text{coarse}}} \mathrm{F1}(\mathcal{S}),$$

where $\mathrm{F1}(\mathcal{S})$ denotes the F1 score achieved using subset $\mathcal{S}$ in a validation setting. This exhaustive subset evaluation yielded 18 non-redundant features that balanced accuracy, interpretability, and clinical relevance (see Table I).

### *C. Model Curation*

Given the binary nature of cardiovascular disease (CVD) outcome and the high clinical cost of both false positives and false negatives, we focused on probabilistic classification methods that support both accurate prediction and interpretable risk estimation. The following models were selected based on their ability to handle the imbalanced, tabular medical dataset:

- **Logistic Regression** – a baseline probabilistic model offering interpretable coefficients and odds ratios, useful for benchmarking.
- **Naïve Bayes** – a generative classifier that assumes conditional independence, often robust in high-dimensional settings.
- **Support Vector Machine (SVM)** – equipped with Platt scaling to produce class probabilities and handle non-linear decision boundaries.
- **Random Forest** – a bagged ensemble that naturally captures variable interactions and ranks feature importance.
- **LightGBM** and **XGBoost** – gradient-boosted frameworks optimized for structured, imbalanced datasets on heterogeneous features.
- **Multilayer Perceptron (MLP)** – a fully connected feed-forward neural network capturing complex non-linear patterns.

Each model was trained using an 80/20 stratified train-test split, preserving CVD class balance across folds. Evaluation focused on two core metrics: F1 score and ROC AUC.

The **F1 score** was prioritized for its balance between precision and recall; critical in CVD screening where both false alarms and missed cases carry clinical consequences. It is defined as the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where $TP$, $FP$, and $FN$ represent true positives, false positives, and false negatives, respectively.

**ROC AUC** was emphasized as a threshold-independent measure of model discrimination, reflecting the model's ability to rank patients by CVD risk across varying decision thresholds. Mathematically, ROC AUC is the area under the Receiver Operating Characteristic curve, which plots the *True Positive Rate* (TPR) against the *False Positive Rate* (FPR):

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

The AUC is then computed as the integral under this curve, typically approximated using the trapezoidal rule.

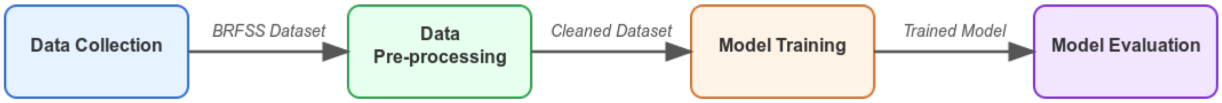

Fig. 1. CDC BRFSS 2023 Data Processing Pipeline

To ensure fair comparison and optimal performance across models, hyperparameter tuning was conducted using randomized search with five-fold cross-validation on the training set. This approach balances exploration of the hyperparameter space with computational efficiency. Key parameters such as the number of estimators (for tree-based models), learning rates (for gradient boosting and MLP), and regularization strengths (for Logistic Regression and SVM) were sampled from predefined distributions. The tuning objective was to maximize the mean cross-validated F1 score, aligning with our focus on minimizing both false positives and false negatives in CVD prediction.

### *D. Interpretability and User Integration*

SHAP (SHapley Additive exPlanations) was initially used for model interpretability. However, it was later replaced with a custom algorithm that manually computes feature contributions and the sigmoid activation function, bypassing `predict_proba` and eliminating the overhead associated with SHAP. This change allows for faster inferences while retaining interpretability.

Feature contributions are defined as:

$$\vec{\phi} = [\phi_1, \ldots, \phi_n],$$

where each $\phi_i$ represents the contribution of feature $i$ to the model's output. The logit is reconstructed by summing these contributions along with the model's bias term $b$:

$$\text{logit} = \sum_{i=1}^{n} \phi_i + b.$$

The predicted probability is then obtained via the sigmoid activation function:

$$P(y = 1) = \frac{1}{1 + e^{-\text{logit}}}.$$

To isolate the most influential predictors, an `argpartition` operation is applied to retrieve the top-$k$ contributing features:

$$\text{top}_k = \text{argpartition}(-\vec{\phi}, k)[: k].$$

The final prediction output, consisting of the estimated risk probability and the corresponding top-$k$ feature attributions, was integrated into an API. This structured representation was then used to train a lightweight large language model (LLM) on annotated CVD and non-CVD profiles derived from the dataset.

Upon receiving some arbitrary input, the model generates two key outputs for users:

1) The predicted risk probability for cardiovascular disease (CVD) based on the user's profile.
2) A set of personalized insights generated by the LLM, identifying the top risk factors contributing to the predicted CVD risk and offering actionable recommendations on how to reduce those risks.

This allows individuals, especially those without clinical expertise to understand not only their predicted risk but also the specific factors influencing their health. The LLM provides feedback focused on actionable advice, allowing users to take informed steps to reduce their CVD risk and consult medical professionals when necessary. Its architecture was intentionally kept simple, given the structured nature of the input and the modest complexity of the output requirements.

## IV. Results

### *A. Model Performance Evaluation*

Seven classification models were evaluated using the preprocessed BRFSS 2023 dataset to predict cardiovascular disease risk, with an 80/20 train/test split. Table II presents the comprehensive performance metrics for each model.

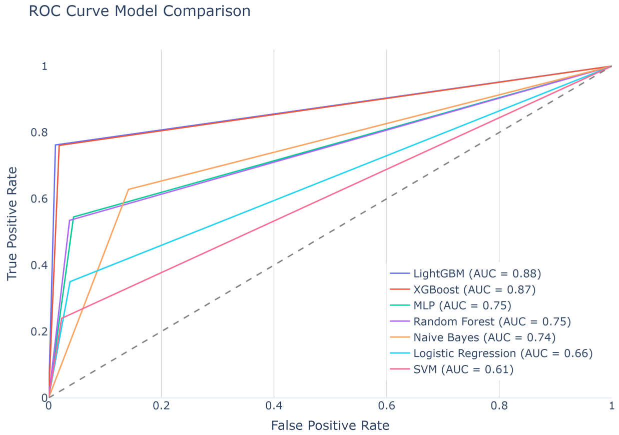

Fig. 2. ROC Curve across all models

Logistic regression, serving as our baseline model, achieved moderate performance with an accuracy of 0.884 and F1 score

TABLE II
PERFORMANCE METRICS OF CVD PREDICTION MODELS

| Model | Accuracy | Precision | Recall | Specificity | F1-Score | ROC AUC |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.884 | 0.573 | 0.350 | 0.962 | 0.435 | 0.656 |
| Naïve Bayes | 0.829 | 0.392 | 0.629 | 0.858 | 0.483 | 0.743 |
| SVM | 0.899 | 0.682 | 0.391 | 0.973 | 0.497 | 0.682 |
| Random Forest | 0.909 | 0.671 | 0.549 | 0.961 | 0.604 | 0.755 |
| **LightGBM** | **0.959** | **0.902** | **0.762** | **0.988** | **0.826** | **0.875** |
| XGBoost | 0.953 | 0.856 | 0.761 | 0.981 | 0.805 | 0.871 |
| Multi-layered Perceptron | 0.904 | 0.642 | 0.545 | 0.956 | 0.590 | 0.751 |

of 0.435. While it demonstrated high specificity (0.962), its relatively low recall (0.350) indicates limited sensitivity in detecting positive CVD cases, a critical concern for clinical screening applications.

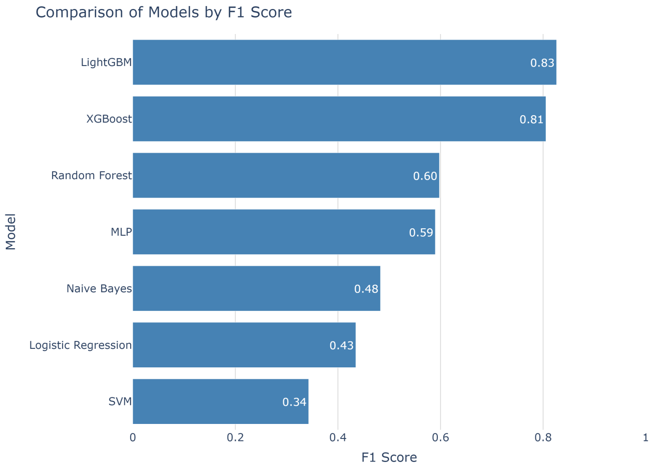

Fig. 3. Ranked F1-scores across models

Naïve Bayes exhibited a different performance profile, with higher recall (0.629) but lower precision (0.392) and specificity (0.858) compared to other models. This indicates a tendency to classify more cases as positive, resulting in increased false positives but fewer missed CVD cases. This trade-off yielded a modest improvement in F1 score (0.483) over the logistic regression baseline.

Support Vector Machine (SVM) with Platt scaling demonstrated improved precision (0.682) over both logistic regression and Naïve Bayes, but its recall (0.391) remained suboptimal. The resulting F1 score (0.497) represented only a marginal improvement over simpler models despite increased computational complexity.

Random Forest showed a more balanced performance profile with substantial improvements in recall (0.549) while maintaining comparable precision (0.671) to SVM. This equilibrium between precision and recall resulted in a notably improved F1 score (0.604), marking it as the best-performing traditional classifier in our evaluation.

### *B. Gradient-Boosted Models Performance*

The gradient-boosted frameworks LightGBM and XGBoost significantly outperformed all other models across all evaluation metrics. LightGBM achieved the highest performance with an F1 score of 0.826, demonstrating high precision (0.902) and recall (0.762) simultaneously. XGBoost followed closely with an F1 score of 0.805, showcasing similar discriminative ability with precision and recall values of 0.856 and 0.761, respectively.

Both gradient-boosted models exhibited superior ROC AUC scores (0.875 for LightGBM and 0.871 for XGBoost), indicating fair discriminative capability across varying classification thresholds. This performance advantage was particularly notable given the class imbalance in the dataset, where the ability to correctly identify positive cases while minimizing false positives is crucial.

### *C. Neural Network Performance*

Despite their theoretical capacity to model complex relationships, neural network approaches underperformed relative to the gradient-boosted models. The implemented Multilayer Perceptron (MLP) achieved an F1 score of 0.590, comparable to Random Forest but significantly below LightGBM and XGBoost. This performance deficit aligns with our expectations regarding neural networks' requirements for large, dense datasets to prevent overfitting.

The MLP's precision (0.642) and recall (0.545) metrics suggest it failed to find an optimal balance between identifying true positives and minimizing false positives. Additional experimentation with more complex architectures, including deeper feedforward networks and Artificial Neural Networks (ANNs), yielded similar or worse performance, with F1 scores consistently in the 0.50-0.60 range.

This may be attributed to the dataset's limited size and sparsity, which likely led to overfitting and poor generalization.

### *D. Model Selection*

Based on comprehensive evaluation, LightGBM was selected as the optimal model for the CVD prediction system. Its superior performance across all metrics, particularly the critical F1 score (0.826) and ROC AUC (0.875), coupled with reasonable computational efficiency, made it well-suited for deployment in both research and potential clinical contexts.

The final deployed system incorporates the LightGBM classifier with the custom feature attribution algorithm to generate personalized risk assessments. These assessments include both a predicted probability of CVD and interpretable explanations of key contributing factors, which are then translated into natural language insights and recommendations by the trained language model.

Post-model development, validation was conducted on a small sample of self-reported data ($n = 25$). The results demonstrated substantial agreement between the model's cardiovascular risk assessments and manual annotations, as indicated by a Cohen's ($\kappa = 0.92$). This supports the model's potential utility for preliminary screening and risk communication applications.

### E. Feature Importance Analysis

The strong performance of tree-based ensemble models, particularly LightGBM, prompted a detailed analysis of feature contributions to CVD prediction. Figure 4 presents the top contributing features identified by the LightGBM model, which was selected for deployment based on its performance.

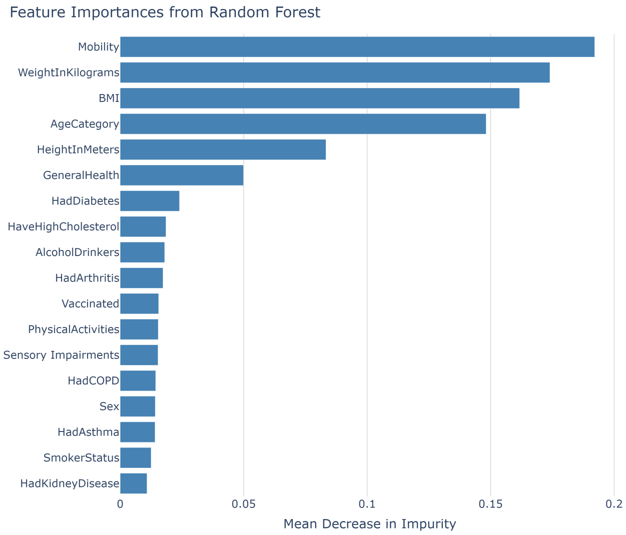

Fig. 4. Feature importances ranked by mean decrease in impurity

Age, high cholesterol, mobility, and BMI were identified as the most influential predictors of cardiovascular disease, consistent with clinical literature. Mobility impairments and the general health status also demonstrated significant predictive value, highlighting the value of lifestyle metrics in cardiovascular disease risk screening.

The custom interpretability algorithm implemented in place of SHAP successfully preserved the granular feature attribution information while significantly reducing computational overhead. When tested on the validation set, the algorithm produced feature contribution rankings highly concordant with traditional importance metrics (Spearman's $\rho = 0.92$), validating its reliability for real-time explanations.

## V. Analysis

### A. Non-Linear Thresholding in Age

Following the identification of feature importance rankings from the LightGBM models, additional analysis was conducted to examine the correlational patterns and predictive behaviors of these prominent variables within the model.

The model exhibited a non-linear threshold effect for certain predictors: incremental changes in these features did not correspond to proportional shifts in predicted cardiovascular disease (CVD) probability until specific values were reached. This effect was most pronounced for the age category variable, aligning with epidemiological evidence of a non-linear increase in CVD incidence with advancing age. As shown in Figure 5, the percentage of CVD cases remains relatively low through the 50–54 age group, then rises sharply beginning at ages 55–59 and continues to increase through the 75–79 category.

Further analysis of the model's handling of the age variable revealed an additional pattern at the upper end of the age distribution. Although the model predicted increasing CVD risk with age through the 75–79 category, predictions for individuals aged 80 and above indicated a plateau or slight decline in CVD probability. This observation, when analyzed, can be attributed to several data-driven explanations based on the properties of the underlying dataset.

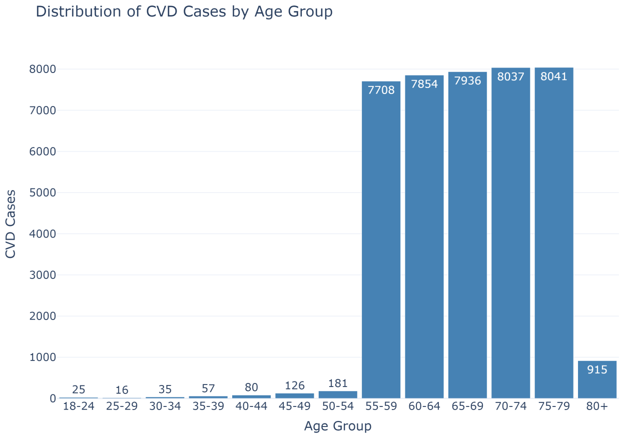

Fig. 5. Distribution of CVD Cases by Age Group

One likely contributing factor is survivorship bias: individuals reaching 80 years or older without prior CVD diagnosis potentially represent a subset characterized by greater cardiovascular resilience or consistent adherence to protective lifestyle factors. Consequently, their probability of developing new CVD at such advanced age shows a relative decrease compared to the risk observed in preceding elderly groups where CVD onset demonstrates peak incidence.

This indicates the model has effectively captured that individuals who have lived eight decades without CVD diagnosis exhibit a different risk profile. Furthermore, this observation may reflect dataset characteristics where individuals developing CVD at younger ages are disproportionately represented compared to new-onset cases in the 80+ demographic, or where competing mortality risks from other conditions become more significant in the very elderly population, thereby influencing the observed CVD incidence rates in the training dataset.

### B. Alcohol Consumption as an Age-Dependent Risk Factor

Complex interactions between lifestyle factors and age, as illustrated in Figure 6, depicts the CVD rate stratified by age category and self-reported drinking status. Consistent with general epidemiological trends and previous observations from this dataset regarding age, CVD rates are minimal in younger age cohorts (e.g., 18-54 years), irrespective of drinking status, with only marginal and somewhat inconsistent differences observed between drinkers and non-drinkers where prevalence is very low. However, a significant increase in CVD prevalence is evident from the 55-59 age category onwards for both groups.

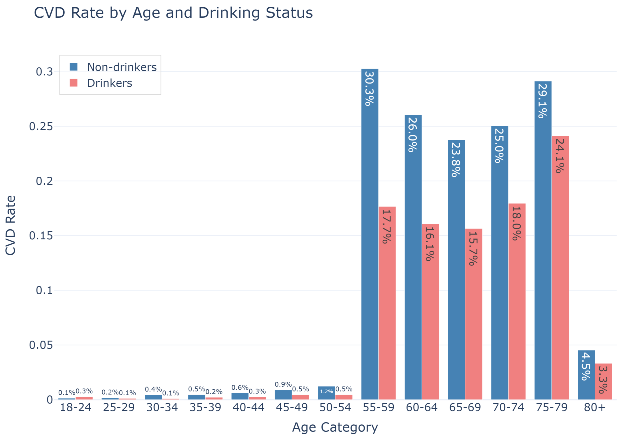

Fig. 6. CVD Rate by Age and Drinking Status

A particularly counterintuitive pattern emerges in these older age groups (specifically 55-79 years): individuals identified as non-drinkers consistently exhibit a markedly higher rate of CVD compared to those identified as drinkers. For instance, in the 55-59 age category, non-drinkers show a 30.3% CVD rate versus 17.7% for drinkers. This disparity, though narrowing slightly, persists across the 60-79 age brackets. This observation, while reflected in the model's learning, warrants careful interpretation.

It may be influenced by factors such as the 'sick quitter' phenomenon, where individuals cease or never began alcohol consumption due to pre-existing health conditions or emerging CVD symptoms, thereby inflating CVD rates within the non-drinker category among older adults. It's also important to note that the 'drinker' category is broad and doesn't differentiate by consumption level (light, moderate, heavy), which has varying health implications. For the 80+ age category, while non-drinkers still show a slightly higher rate (4.5% vs. 3.3%), the overall CVD rates for both groups decline sharply from the preceding age groups, further supporting the potential survivorship effects discussed earlier. Consequently, the model, when trained on this data, would likely learn to associate non-drinking status with an increased probability of CVD, particularly in older populations (55-79), while potentially assigning less predictive weight to drinking status in younger individuals or the very elderly (80+) where other factors or data characteristics dominate.

### C. Smoking Status as an Age-Dependent Risk Factor

The influence of smoking status on CVD rates across different age categories was also prominently evident in the dataset, as depicted in Figure 7. This figure illustrates a consistent and significant pattern: individuals identified as smokers exhibit a higher prevalence of CVD compared to non-smokers across nearly all age groups. While CVD rates are generally low in younger cohorts (18-49 years), smokers in these groups still consistently show slightly elevated rates, for instance, 0.7% vs. 0.2% for smokers and non-smokers respectively in the 18-24 age category. The disparity becomes increasingly pronounced with advancing age. Starting from the 50-54 age group, the CVD rate for smokers (1.7%) is notably higher than for non-smokers (0.6%), and this gap widens considerably thereafter. For example, in the 55-59 age category, smokers have a 27.0% CVD rate compared to 23.0% for non-smokers. This trend continues, culminating in a dramatic difference in the 75-79 age group, where smokers demonstrate a 48.7% CVD rate, nearly double that of non-smokers (24.8%). Even in the 80+ age category, where overall rates are lower likely due to survivorship, smokers still present a higher CVD rate (5.0%) than non-smokers (4.0%). This strong and consistent association within the dataset would undoubtedly lead the model to learn smoking status as a significant risk factor for CVD, increasing the predicted probability for individuals who smoke, particularly as age increases.

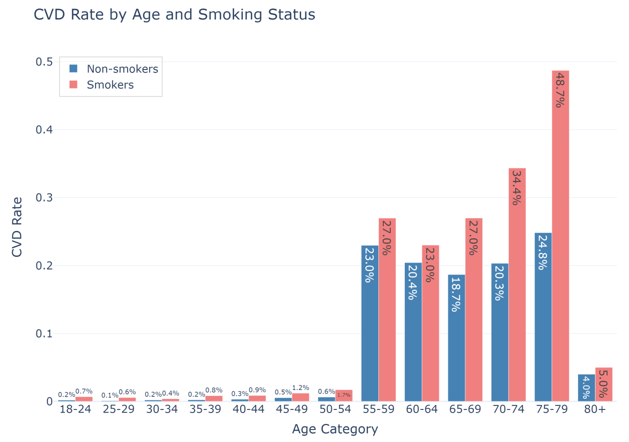

Fig. 7. CVD Rate by Age and Smoking Status

### D. Threshold Effects in BMI and Obesity-Linked Risk

A similar pattern of threshold effects, where risk escalates more sharply beyond certain points, was discernible for Body Mass Index (BMI), another critical predictor variable. Figure 8 presents the BMI density curves for individuals with and without Cardiovascular Disease (CVD), offering a visual representation of how the model learns to associate different BMI levels with CVD risk based on the underlying data distributions.

Observing the teal line, which represents individuals without CVD (”No CVD”), it is evident that this group's BMI distribution peaks within the range typically considered healthy to slightly overweight, approximately at a BMI of 25-27. This indicates that a high concentration of individuals without CVD have BMIs in this spectrum. Beyond this peak, the probability density for the non-CVD group declines sharply, becoming considerably lower as BMI values increase, particularly as they enter the obese categories (BMI $>$ 30). This sharp decline signifies that very high BMIs are less common among individuals without CVD in the dataset.

In contrast, the red line, representing individuals with CVD, shows a distribution that is markedly shifted towards higher BMI values. The peak of the CVD group's density curve occurs at a higher BMI, around 28-30, which falls into the overweight to early obese (Class I) range. While the model would learn a relatively modest increase in predicted CVD probability as BMI transitions from the ideal range (e.g., 18.5-24.9) through the overweight range (e.g., 25-29.9), a more significant pattern emerges when BMI values cross into obesity.

Specifically, for BMI values exceeding approximately 30, the red curve (CVD group) consistently maintains a higher probability density, or declines much more slowly, compared to the teal curve (non-CVD group). For instance, in the broad BMI range of 30 to well over 40, the red curve is distinctly above the teal curve, or the teal curve has diminished to near zero while the red curve still shows a notable presence. This visual disparity underscores that a significantly greater proportion of individuals in the obese categories are found within the CVD-positive cohort compared to the non-CVD cohort.

This difference in distributions is what the model learns. It interprets the higher concentration of CVD cases at these elevated BMI levels as a strong signal for increased CVD probability. The model's predicted likelihood of CVD, therefore, demonstrates a more pronounced escalation once BMI values move significantly into the obese categories. The clear separation and differing shapes of the two curves, particularly the rightward skew and the ”fatter tail” of the CVD curve extending into higher BMIs, effectively train the model to recognize that while moderate increases in BMI have some impact, crossing the threshold into obesity substantially amplifies the predicted cardiovascular risk.

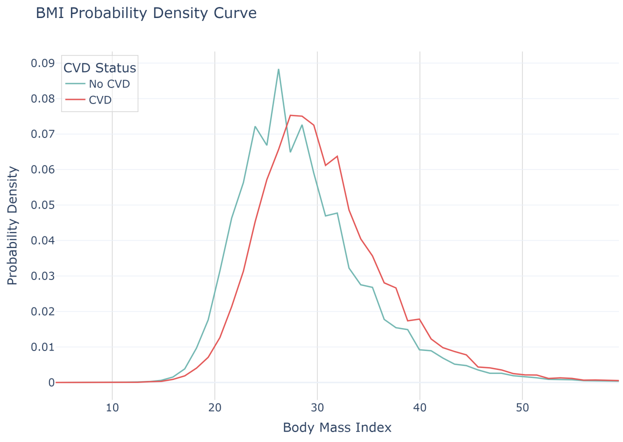

Fig. 8. BMI Probability Density Curve

### E. Compounding Features

In addition to thresholding, another noteworthy characteristic was the synergistic effect observed between specific pairs of features. Certain variables, while exhibiting only modest influence on CVD probability when assessed independently, demonstrated significantly enhanced predictive power when considered in conjunction with other relevant factors. This was particularly evident in the interplay between mobility and age, where the combination of advanced age and reduced mobility was a more potent predictor of CVD than the additive effects of each variable alone would suggest.

In essence, the preceding analysis demonstrates that the developed model operates as a data-driven instrument, meticulously learning and reflecting the statistical patterns, thresholds, and interactive effects present within the BRFSS dataset. The observed predictive behaviors concerning age, BMI, drinking, and smoking status, while often aligning with established clinical knowledge, are fundamentally derived from these learned associations in the data rather than from an embedded clinical inference engine. This inherent reliance on the input data necessitates a critical examination of the various methodological constraints and data characteristics that could influence its performance and interpretation.

## VI. Limitations

### A. Data Quality and Availability

Several methodological constraints should be considered when interpreting the results of this study. The availability of high-quality medical data posed a challenge, resulting in the use of BRFSS lifestyle data instead of clinically validated measurements. Although the BRFSS dataset provides broad population coverage, its periodic reliance on self-reported data introduces potential reporting biases that may affect the accuracy of both predictor and outcome variables.

### B. Null Data Handling

The dataset reduction from 433,016 to 322,789 instances due to missing values represents a substantial loss of potentially informative data points. This complete-case analysis approach, while methodologically straightforward, may have introduced selection bias if the missing data mechanism was not completely random. Alternative imputation strategies could potentially have preserved more observations while maintaining data integrity.

### C. Class Imbalance

Class imbalance was another constraint. Synthetic augmentation techniques, including SMOTE (Synthetic Minority Over-sampling Technique), were considered but not implemented. This decision was based on concerns regarding the introduction of artificial patterns that could reduce model generalizability. Additionally, empirical testing of SMOTE yielded no performance improvement.

### D. Data Inconsistency

Data consistency issues also constrained the analysis. Recent Trump policy changes affecting the BRFSS survey led to the removal of certain variables across survey years, complicating longitudinal analysis and cross-year data integration. Specifically, features such as sleep patterns and detailed race/ethnicity information were inconsistently available, limiting our ability to incorporate these potentially relevant predictors into the models.

### E. Computational Constraints

Hyperparameter optimization was limited by computational resources. Although a systematic tuning approach was applied, it did not exhaustively explore the parameter space. More advanced techniques such as Bayesian optimization or genetic algorithms were not feasible under the given constraints.

### F. Validation Limitations

The validation strategy also presents limitations. Post-development validation on a self-reported sample (n = 25) was insufficient for robust external validation. The volunteer-based sample may have introduced selection bias, as participants could differ systematically from the general population with respect to cardiovascular risk. Furthermore, no clinical verification of reported cardiovascular conditions was conducted.

### G. Model Architecture Constraints

Additionally, the model development did not incorporate more advanced deep learning architectures beyond standard MLPs, primarily due to their poor performance on the available structured data. While state-of-the-art transformer models have shown promise in certain healthcare applications, their requirement for larger training datasets and their computational complexity rendered them impractical for this specific implementation.

### H. Literature Comparison Challenges

The research was further complicated by data provenance issues in the cardiovascular prediction literature. Many studies reference datasets of unclear origin or with limited documentation of preprocessing steps. For instance, the frequently cited ”Cleveland” heart disease dataset appears in numerous publications but with inconsistent descriptions and accessibility, creating challenges for direct performance comparisons across studies.

### I. Generalizability Concerns

Finally, the single-institution nature of the post-development validation limits generalizability. A multi-center validation approach involving diverse healthcare settings and patient populations would provide more robust evidence of the model's external validity and clinical utility across different contexts.

## VII. Future Work

This research presents many directions for future development and enhancement.

### A. Advanced Data Handling Techniques

The development of more sophisticated methods for handling missing data represents a promising direction for improvement. Standard imputation techniques were insufficient for preserving data quality in the current analysis, resulting in the exclusion of 110,227 instances. Implementing advanced approaches such as multiple imputation or deep learning-based imputation models could allow for the retention of these cases, thereby increasing statistical power and potentially improving model generalizability.

### B. Feature Engineering and Representation Learning

The investigation of alternative feature engineering approaches constitutes another area for future work. The current study utilized domain knowledge and statistical filtering to select features, but automated feature extraction methods such as autoencoders or representation learning could identify latent patterns not captured by traditional approaches. Additionally, exploration of non-linear feature interactions through techniques such as automatic feature crossing may further enhance predictive performance.

### C. External Validation and Generalizability

External validation represents an important next step. Future work should include multi-center validation studies across diverse healthcare settings and demographic populations to assess the model's generalizability beyond the initial development. Prospective validation would be particularly valuable, involving the collection of longitudinal data to evaluate the model's predictive performance for incident cardiovascular disease rather than prevalent cases.

### D. Multimodal Data Integration

The integration of additional data modalities presents another area of improvement. Combining the BRFSS lifestyle data with electronic health record (EHR) data, genomic information, or wearable device metrics could potentially provide a more comprehensive view of cardiovascular risk. This multimodal approach may capture risk factors beyond those available in the current dataset.

### E. Advanced Neural Network Architectures

Development of more advanced deep learning architectures specifically designed for tabular medical data represents another potential extension. While standard MLP approaches underperformed in the current analysis, recent advances in transformer-based models for tabular data may offer improved performance when sufficient training data is available. Exploration of specialized architectures that effectively handle mixed continuous and categorical variables could address the limitations of traditional neural networks in this domain.

### F. Longitudinal Risk Progression Modeling

Investigation of time-series modeling approaches could also enhance the current work. Utilizing longitudinal BRFSS data across multiple years would allow for the development of models that capture the progression of cardiovascular risk over time, potentially offering improved predictive performance compared to static approaches. This would require addressing the challenge of variable inconsistency across survey years.

### G. Fairness and Algorithmic Bias Mitigation

The calibration and fairness of prediction models across demographic subgroups warrants further investigation. Future work should systematically assess model performance across race, ethnicity, gender, and socioeconomic strata to identify and mitigate potential biases. The development of fairness-aware learning algorithms could potentially address disparities in predictive performance across population subgroups.

### H. Clinical Implementation and Decision Support

The translation of predictive models into clinical decision support tools represents an important direction for future research. This would involve the development of user-centered interfaces that effectively communicate risk information to healthcare providers and patients, integration with existing clinical workflows, and evaluation of the impact on clinical decision-making and patient outcomes.

## VIII. Conclusion

This study developed and evaluated several probabilistic classification models for cardiovascular disease risk prediction using the CDC's Behavioral Risk Factor Surveillance System dataset. Through a systematic feature selection process that incorporated clinical domain knowledge and statistical filtering, 18 non-redundant predictors were identified from an initial pool of 350 candidate variables. The gradient-boosted models, particularly LightGBM, demonstrated superior predictive performance with an F1 score of 0.826 and ROC AUC of 0.875, significantly outperforming traditional classifiers and neural network approaches.

The performance differential between model architectures underscores the importance of selecting appropriate algorithmic frameworks for medical risk prediction tasks. While neural networks often excel in domains with large, dense datasets, our results indicate that gradient-boosted decision trees offer superior performance for structured, heterogeneous healthcare data with moderate class imbalance. This finding aligns with prior research in clinical prediction modeling where tree-based ensembles have consistently shown robust performance on tabular data.

Feature importance analysis revealed age, high cholesterol, mobility impairments, and BMI as the most influential predictors of cardiovascular disease risk, consistent with established clinical literature. The identification of mobility impairments and general health metrics as significant predictors highlights the potential utility of incorporating functional status assessments into cardiovascular screening protocols. The custom interpretability algorithm implemented in place of SHAP successfully preserved feature attribution information while reducing computational overhead, enabling real-time explanations of model predictions.

The integration of the LightGBM classifier with a language model for generating personalized risk insights represents a novel approach to clinical decision support that balances predictive accuracy with interpretability. This architecture allows for the translation of complex statistical outputs into actionable recommendations, potentially improving risk communication in clinical and public health settings. Post-development validation on self-reported data demonstrated substantial agreement between model predictions and manual annotations Cohen's ($\kappa = 0.92$), supporting the system's potential utility for preliminary screening applications.

Future work should focus on longitudinal validation using electronic health record data to assess the model's predictive stability over time and across diverse patient populations. Additionally, exploration of hybrid modeling approaches that combine structured clinical data with unstructured information from clinical notes may further enhance predictive accuracy. The development of calibrated risk thresholds specific to different demographic subgroups represents another important direction for improving the clinical utility of the prediction system.

In conclusion, this study demonstrates the feasibility of developing accurate, interpretable cardiovascular disease prediction models using population-level survey data. The superior performance of gradient-boosted models, coupled with transparent feature attribution mechanisms, offers promising avenues for enhancing cardiovascular risk assessment in both clinical and public health contexts. The methodological framework presented may serve as a template for developing similar risk prediction tools for other chronic conditions, potentially contributing to more efficient resource allocation and targeted preventive interventions.

## Acknowledgment

We express our sincere gratitude to the Department of Computing and Information Technology (DCIT) at The University of the West Indies for fostering an academic environment that enabled the successful completion of this research. The resources, infrastructure, and institutional support provided by the department played a critical role in facilitating our exploration and development throughout the duration of this project.

We are especially thankful to M.Sc. Sergio Mathurin, whose leadership, vision, and mentorship were foundational to the progress and success of our work. From designing the initial course framework to offering ongoing feedback, encouragement, and insightful critique, his dedication ensured that we remained focused and motivated. His willingness to share his deep expertise and to invest time in our learning and development significantly elevated the quality of this project. His guidance not only impacted our own research but also enhanced the overall learning experience of our cohort.

Additionally, we would like to acknowledge the broader DCIT staff, whose tireless efforts behind the scenes ensured the seamless execution of the course. Their professionalism, timely communication, and logistical coordination contributed immensely to the structure and smooth progression of the research process. We are grateful for their unwavering support, and we recognize the importance of their collective contribution to our academic and professional growth.

Finally, we extend our appreciation to our peers and colleagues whose collaboration, discussions, and feedback helped refine our ideas and approaches, fostering a stimulating environment of shared learning.

## References

[1] K. Pluta et al., “Platelet–Leucocyte aggregates as novel biomarkers in cardiovascular diseases,” Biology, vol. 11, no. 2, p. 224, Jan. 2022, doi: 10.3390/biology11020224.

[2] O. Taylan, A. Alkabaa, H. Alqabbaa, E. Pamukçu, and V. Leiva, “Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods,” Biology, vol. 12, no. 1, p. 117, Jan. 2023, doi: 10.3390/biology12010117.

[3] S. Subramani et al., “Cardiovascular diseases prediction by machine learning incorporation with deep learning,” Frontiers in Medicine, vol. 10, Apr. 2023, doi: 10.3389/fmed.2023.1150933.

[4] F. Bray et al., “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA a Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, Apr. 2024, doi: 10.3322/caac.21834.

[5] “Global cancer facts & figures,” American Cancer Society. https://www.cancer.org/research/cancer-facts-statistics/global-cancer-facts-and-figures.html

[6] S. C. Curtin M. A., B. Tejada-Vera M. S., and B. A. Bastian B. S., “Deaths: Leading Causes for 2022,” National Vital Statistics Reports, vol. 73, no. 10, Dec. 2024, [Online]. Available: https://www.cdc.gov/nchs/data/nvsr/nvsr73/nvsr73-10.pdf

[7] S. S. Martin et al., “2025 Heart Disease and Stroke Statistics: A report of US and global data from the American Heart Association,” Circulation, Jan. 2025, doi: 10.1161/cir.0000000000001303.

[8] World Health Organization: WHO, “Cardiovascular diseases (CVDs),” Jun. 11, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[9] “Health effects of cigarettes: cardiovascular disease,” Smoking and Tobacco Use, Sep. 17, 2024. https://www.cdc.gov/tobacco/about/cigarettes-and-cardiovascular-disease.html

[10] E. Jung, S. Y. Kong, Y. S. Ro, H. H. Ryu, and S. D. Shin, “Serum Cholesterol Levels and Risk of Cardiovascular Death: A Systematic Review and A Dose-Response Meta-Analysis of Prospective Cohort Studies,” International Journal of Environmental Research and Public Health, vol. 19, no. 14, p. 8272, Jul. 2022, doi: 10.3390/ijerph19148272.

[11] M. C. Bertoluci and V. Z. Rocha, “Cardiovascular risk assessment in patients with diabetes,” Diabetology & Metabolic Syndrome, vol. 9, no. 1, Apr. 2017, doi: 10.1186/s13098-017-0225-1.

[12] R. R. Baiduc, J. W. Sun, C. M. Berry, M. Anderson, and E. A. Vance, “Relationship of cardiovascular disease risk and hearing loss in a clinical population,” Scientific Reports, vol. 13, no. 1, Jan. 2023, doi: 10.1038/s41598-023-28599-9.

[13] I. Mendez, M. Kim, E. A. Lundeen, F. Loustalot, J. Fang, and J. Saaddine, “Cardiovascular disease risk factors in US adults with vision impairment,” Preventing Chronic Disease, vol. 19, Jul. 2022, doi: 10.5888/pcd19.220027.

[14] G. Lippi, B. M. Henry, and F. Sanchis-Gomar, “Physical inactivity and cardiovascular disease at the time of coronavirus disease 2019 (COVID-19),” European Journal of Preventive Cardiology, vol. 27, no. 9, pp. 906–908, Apr. 2020, doi: 10.1177/2047487320916823.

[15] M. L. Wilby, “Physical mobility impairment and risk for cardiovascular disease,” Health Equity, vol. 3, no. 1, pp. 527–531, Oct. 2019, doi: 10.1089/heq.2019.0065.

[16] S. García-Zamora and L. Pulido, “Vaccines in cardiology, an underutilized strategy to reduce the residual cardiovascular risk,” Archivos Peruanos De Cardiología Y Cirugía Cardiovascular, vol. 5, no. 1, pp. 29–39, Mar. 2024, doi: 10.47487/apcyccv.v5i1.349.

[17] BRFSS. Accessed: May. 10, 2025. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2023.html

[18] M. Siddhartha, “Heart Disease Dataset (Comprehensive).” IEEE Dataport, Nov. 05, 2020. doi: 10.21227/dz4t-cm36.

[19] B. P. Doppala and D. Bhattacharyya, “Cardiovascular_Disease_Dataset.” Mendeley Data, Apr. 16, 2021. doi: 10.17632/dzz48mvjht.1.

[20] “Heart disease dataset,” Kaggle. Jun. 06, 2019. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[21] “Cardiovascular Disease dataset,” Kaggle. Jan. 20, 2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[22] “Heart Disease Prediction Dataset,” Kaggle. Sep. 27, 2024. [Online]. Available: https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset

[23] S. Sreekumari, R. Bhalla, and G. Ganesan, “A Comparative Study of Heart Disease Prediction using Machine Learning,” in Joint Proceedings of the Workshop on Intelligent Systems (WINS 2023) and the Workshop on Computer Vision and Machine Learning for Healthcare (CVMLH 2023), Chennai, India, May. 2023, pp. 54–65. Available: https://ceur-ws.org/Vol-3635/ICCS_CVMLH_01.pdf

[24] N. Rahman, Md. A. Mahbub, and Md. H. Mahbub, “Enhancing Heart Disease Diagnosis with Machine Learning Algorithms: An Evaluation on Imbalanced Data and Performance Metrics,” 2024 2nd International

Conference on Information and Communication Technology (ICICT), pp. 269–273, Oct. 2024, doi: 10.1109/icict64387.2024.10839651.

[25] Tink-a-Ton, “Cardiovascular Disease Prediction,” GitHub. https://github.com/carrot2803/cvd-predictor/ (accessed May 09, 2025).

## Appendix

### Appendix A: Prediction Model Implementation

The cardiovascular disease prediction model is implemented across the following repositories:

**Backend Repository**: Implements data preprocessing, feature engineering, model training, and inference logic. `https://github.com/carrot2803/cvd-predictor`

**Frontend Repository**: Provides an interface for survey-based input and displays tailored feedback from the LLM. `https://github.com/carrot2803/cvd-app`

**Auxiliary Materials**: Provides the materials needed to run the codebase, along with a summary of challenges faced, etc. `https://shorturl.at/ncjVr`