The Effects of Sleep, Stress, and Lifestyle on Heart Disease: Evidence from BRFSS Data

Kerem Kuru (220303016)
Kayra Ege Kosaoglu (220303039)
Baris Gundogdu (230303912)
Department of Computer Engineering
Istanbul Arel University
Istanbul, Turkey

***Abstract*—Heart disease is a general term for conditions that affect the structure or function of the heart and disrupt blood circulation. These conditions include coronary artery disease, heart failure, arrhythmias, heart valve disorders, and congenital heart diseases. They are often associated with factors such as genetic predisposition, unhealthy diet, smoking, high blood pressure, diabetes, and stress.**

Unrecognized heart disease may not show obvious symptoms, especially in the early stages, and individuals often become aware of the condition during routine health examinations or when a serious problem occurs.

This study aims to predict the likelihood of developing heart disease based on daily lifestyle habits, genetic factors, and social conditions. Logistic Regression, Random Forest, and XGBoost models are used to achieve accuracy scores, while the research additionally includes original and novelty-based approaches.

***Index Terms*—Heart disease, machine learning, BRFSS, class imbalance, logistic regression, XGBoost, lifestyle analysis.**

I. Introduction

Heart disease is a general term for disorders that affect the structure or function of the heart and disrupt blood circulation. These conditions include coronary artery disease, heart failure, arrhythmias, heart valve disorders, and congenital heart diseases [1], [2].

Unrecognized heart disease may not present clear symptoms, especially in its early stages, and individuals often become aware of the condition during routine medical checkups or when a serious complication arises [3]. In men, heart attack symptoms typically include severe chest pain, pain radiating to the left arm or jaw, and shortness of breath.

Women may experience similar symptoms; however, their pain can be more widespread, affecting the shoulders, neck, arms, abdomen, and back. In addition, pain in women may resemble indigestion and may not be persistent. A heart attack may also occur without pain, but with unexplained anxiety, nausea, dizziness, palpitations, and cold sweating [4]. Unexplained fatigue may appear in women before a heart attack.

Cardiovascular diseases, including heart disease, are easier to treat when detected early [2]. These diseases can be examined across different categories based on factors such as gender, age, medical history, and social lifestyle [5], [6]. Advancements in data analysis methods within biotechnology and biology have led to the emergence of bioinformatics, while machine learning techniques have enabled scalable and effective predictive modeling approaches [7], [8].

In this study, novel interpretations and modeling strategies are introduced by utilizing a large-scale cardiovascular risk dataset. The primary objective is to explore the relationship between lifestyle behaviors, mental well-being, and medical history in the prediction of heart disease.

II. Dataset and Feature Groups

A. Dataset Description

The dataset used in this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS), a large-scale health-related survey conducted annually in the United States by the Centers for Disease Control and Prevention [9]. The dataset consists of self-reported information on lifestyle behaviors, chronic health conditions, and demographic characteristics.

Following preprocessing, the final dataset contains 319,795 instances with 18 attributes, including a binary target variable indicating the presence of heart disease. Similar to many real-world healthcare datasets, the class distribution is highly imbalanced, with a substantially lower number of positive heart disease cases compared to negative cases [10].

B. Feature Grouping Strategy

To systematically evaluate the contribution of different predictors, features were grouped based on their semantic meaning and clinical relevance. This grouping strategy aligns with previous cardiovascular risk modeling studies [5], [8].

1) Lifestyle Features: Lifestyle-related variables include smoking status, alcohol consumption, physical activity, and body mass index (BMI). These factors are widely recognized as modifiable determinants of cardiovascular health and have been strongly associated with increased heart disease risk [6], [11].

2) Sleep and Stress Features: Sleep duration and mental health indicators were incorporated to capture the effects of rest quality and psychological stress. Insufficient sleep and chronic stress have been linked to increased cardiovascular morbidity and mortality [12], [13].

3) Demographic Features: Demographic attributes such as age and sex were included to account for population-level differences in cardiovascular disease prevalence. These variables are commonly used in clinical risk assessment models [4], [5].

4) Medical History Features: Medical history variables include prior stroke events, diabetes, kidney disease, and other chronic conditions. Such clinical indicators are known to significantly increase cardiovascular risk and are frequently incorporated into predictive risk models [3], [6].

C. Model Configurations

Based on the defined feature groups, two different model configurations were constructed:

- **Model A (Lifestyle-Oriented Model):** Includes lifestyle, sleep, stress, and demographic features.
- **Model B (Extended Clinical Model):** Includes all features from Model A with the addition of medical history variables.

This structured grouping enables a comparative analysis of predictive performance and forms the basis of the proposed novelty methodology, where the impact of medical history inclusion is evaluated under balanced learning conditions.

III. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine the relationships between lifestyle, mental health, and demographic variables with heart disease prevalence. Previous studies have emphasized the importance of exploratory analysis in uncovering meaningful patterns within complex biomedical datasets [7], [8].

Observed trends indicate that insufficient sleep, reduced physical activity, higher BMI, and increased psychological stress are associated with elevated heart disease prevalence, consistent with prior epidemiological findings [11]–[13].

A. Sleep Duration

Sleep duration plays a critical role in cardiovascular regulation. Individuals with insufficient or excessive sleep may experience hormonal imbalance, increased inflammation, and metabolic dysregulation, all of which can contribute to elevated heart disease risk. The distribution below suggests that both very short and prolonged sleep durations are associated with higher heart disease prevalence compared to moderate sleep durations.

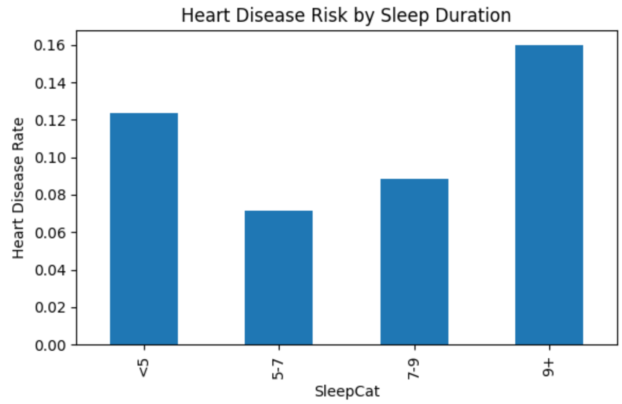

Fig. 1. Heart disease prevalence across sleep duration categories.

B. Mental Health

Mental health has increasingly been recognized as an important factor in cardiovascular outcomes. Chronic psychological stress and prolonged periods of poor mental health may influence cardiovascular risk through behavioral changes and physiological stress responses. The figure below indicates a gradual increase in heart disease prevalence as the number of poor mental health days increases.

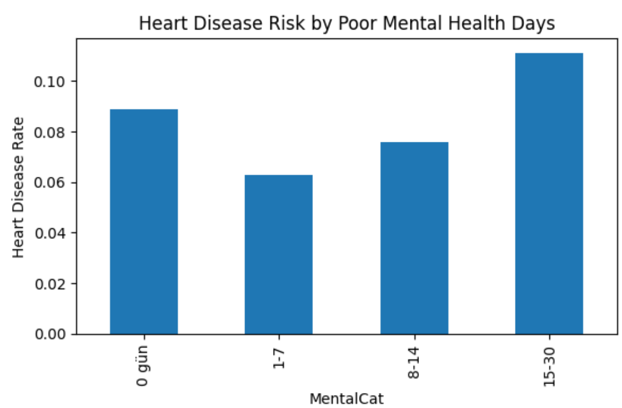

Fig. 2. Heart disease prevalence by poor mental health days.

C. Smoking

Smoking is one of the most established modifiable risk factors for cardiovascular disease. Long-term exposure to smoking leads to endothelial dysfunction, increased oxidative stress, and reduced oxygen transport capacity. As shown below, individuals who report smoking exhibit substantially higher heart disease prevalence compared to non-smokers.

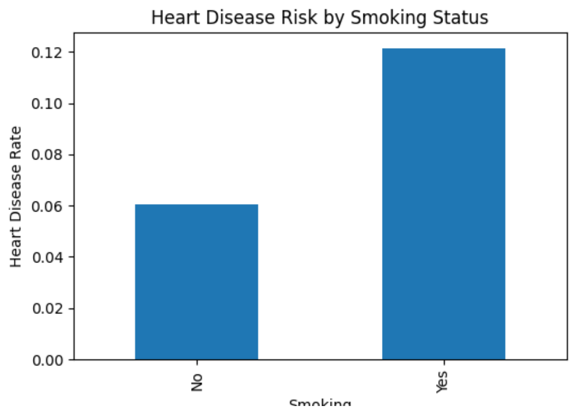

Fig. 3. Heart disease prevalence by smoking status.

D. Alcohol Consumption

Alcohol consumption shows complex associations with cardiovascular health depending on intake frequency and quantity. The figure below presents heart disease prevalence by reported alcohol drinking status. The observed difference between groups should be interpreted cautiously, as alcohol consumption may interact with other lifestyle and demographic factors.

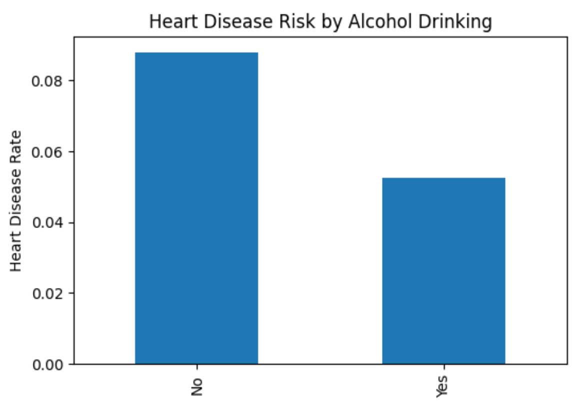

Fig. 4. Heart disease prevalence by alcohol drinking status.

E. Physical Activity

Regular physical activity is commonly associated with improved cardiovascular health through enhanced metabolic efficiency and reduced inflammation. The figure below demonstrates that individuals who report engaging in physical activity have a lower prevalence of heart disease compared to physically inactive individuals.

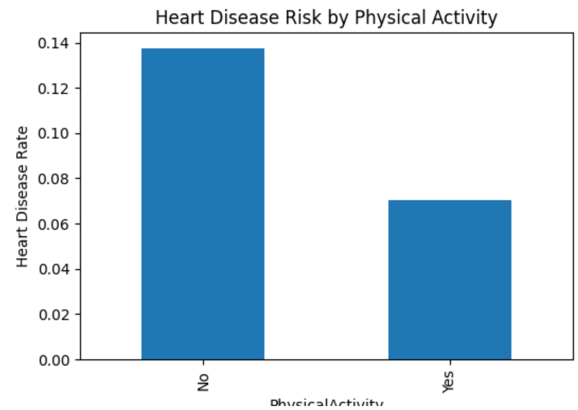

Fig. 5. Heart disease prevalence by physical activity status.

F. Body Mass Index (BMI)

Body Mass Index (BMI) serves as an indicator of obesity-related cardiovascular risk. Increased BMI is often linked to hypertension, insulin resistance, and systemic inflammation. As illustrated below, higher BMI categories correspond to increased heart disease prevalence.

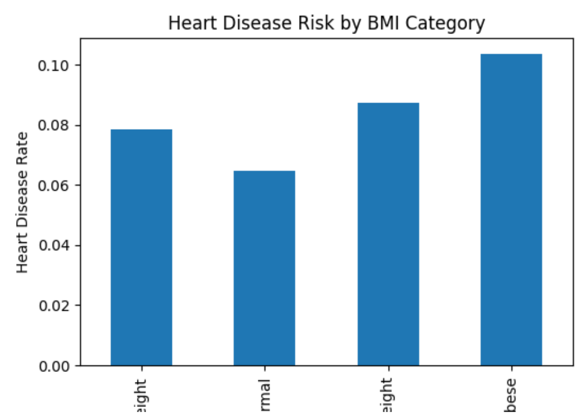

Fig. 6. Heart disease prevalence by BMI category.

G. Age

Age is a major non-modifiable risk factor for cardiovascular disease. The prevalence of heart disease increases progressively with age, reflecting cumulative exposure to risk factors and physiological aging processes. The following figure demonstrates a clear upward trend in heart disease prevalence across age categories.

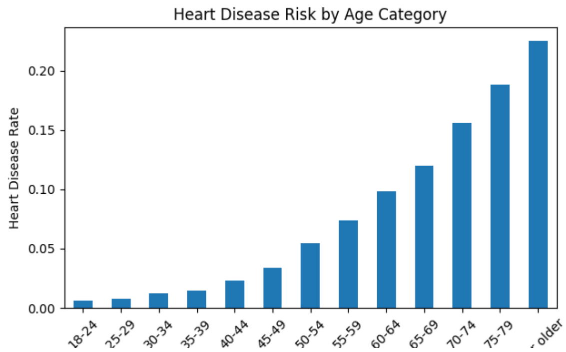

Fig. 7. Heart disease prevalence by age category.

Overall, the EDA findings highlight meaningful associations between heart disease prevalence and key lifestyle, mental health, and demographic variables. These observations provide empirical justification for the feature grouping strategy and the balanced learning-based novelty methodology employed in this study.

IV. Methodology

A. Dataset Description and Preprocessing

The study utilizes a cleaned version of the Behavioral Risk Factor Surveillance System (BRFSS) dataset, consisting of 319,795 records with 18 attributes. The target variable, *HeartDisease*, is a binary indicator representing whether an individual has been diagnosed with a heart-related condition.

Similar to many real-world healthcare datasets, the class distribution is highly imbalanced, with significantly fewer positive heart disease cases compared to negative cases [14]. Missing values were handled by removing incomplete records within selected feature subsets. Categorical variables were transformed using one-hot encoding, while numerical features were standardized where required by the applied learning algorithms.

B. Feature Grouping Strategy

To systematically evaluate the contribution of different predictors, input features were grouped according to their semantic meaning and clinical relevance, following common practices in cardiovascular risk modeling studies [15].

1) Lifestyle Features: Lifestyle-related variables represent daily habits that may influence cardiovascular health. This group includes smoking status, alcohol consumption, physical activity, body mass index (BMI), and general health perception. Prior studies have strongly associated these factors with increased cardiovascular risk [16].

2) Sleep and Stress Features: Sleep duration and stress-related indicators were incorporated to capture the effects of rest quality and mental well-being on heart health. Variables such as sleep time, mental health days, and physical health days were used to represent chronic stress exposure. Insufficient sleep and prolonged stress have been identified as important contributors to cardiovascular dysfunction [17], [18].

3) Demographic Features: Demographic information includes age category, sex, and race. These variables are commonly used in cardiovascular risk prediction models to account for population-level differences in disease prevalence [19].

4) Medical History Features: Medical history features represent pre-existing health conditions that may directly increase cardiovascular risk. This group includes prior stroke events, difficulty in walking, diabetes, asthma, kidney disease, and skin cancer. Such conditions are widely recognized as strong clinical indicators of cardiovascular vulnerability [20].

V. Model Configurations

Three supervised machine learning models were initially evaluated: Logistic Regression, Random Forest, and XGBoost. Logistic Regression was selected due to its interpretability and widespread use in medical decision-making. Random Forest was employed as a tree-based ensemble method capable of capturing nonlinear relationships, while XGBoost was chosen for its strong performance in structured healthcare datasets [21].

All models were trained using an 80/20 train-test split with stratified sampling to preserve class distribution. Standard performance metrics included accuracy, ROC-AUC, precision, recall, and F1-score. Given the clinical objective of identifying individuals at risk of heart disease, recall for the positive class was considered a critical evaluation metric.

VI. Novelty Methodology

A. Partial Novelty: Impact of Class Imbalance on Model Performance

In healthcare prediction tasks, class imbalance poses a significant challenge, often leading models to favor the majority class and underestimate rare but clinically important outcomes [22]. To address this issue, the performance of Logistic Regression, Random Forest, and XGBoost models was evaluated under both unbalanced and balanced learning settings.

In the unbalanced configuration, all three models achieved high accuracy and ROC-AUC values; however, recall for the positive heart disease class remained notably low. For instance, unbalanced Logistic Regression and XGBoost models demonstrated recall values below 0.06, indicating poor sensitivity in detecting heart disease cases despite favorable accuracy scores.

To mitigate this limitation, class-weighted learning was applied. Logistic Regression utilized balanced class weights, while XGBoost incorporated a scale-positive-weight parameter derived from class frequency ratios. Random Forest was also evaluated using balanced class weights for comparison purposes.

The balanced models exhibited a substantial improvement in positive-class recall, increasing from below 0.06 to approximately 0.78 for both Logistic Regression and XGBoost. Although this improvement was accompanied by a decrease in overall accuracy, the trade-off is considered acceptable in medical risk prediction contexts, where minimizing false negatives is often prioritized [23].

B. Novelty: Feature-Group–Driven Prediction Strategy

The primary novelty of this study lies in the comparative evaluation of feature-group–driven models under balanced learning conditions. Two model configurations were constructed to assess the incremental contribution of medical history information.

Model A (Lifestyle-Oriented Model) includes lifestyle, sleep, stress, and demographic features. **Model B (Extended Clinical Model)** extends Model A by incorporating medical history features.

Both configurations were evaluated using balanced Logistic Regression and XGBoost models. Random Forest was excluded from this novelty analysis due to its consistently low recall values for the positive class, which limits its clinical applicability in heart disease detection tasks.

Experimental results demonstrate that Model B outperforms Model A across all evaluated metrics. The inclusion of medical history features led to measurable improvements in accuracy, ROC-AUC, precision, and F1-score, while maintaining high recall values above 0.77 for both models. This indicates that medical history provides complementary predictive information beyond lifestyle and demographic factors alone.

These findings suggest that combining lifestyle-driven indicators with clinically grounded medical history features enhances heart disease prediction performance under balanced learning conditions, thereby supporting the proposed novelty approach.

C. Evaluation Metrics

Model performance was evaluated using Accuracy, ROC-AUC, Precision, Recall, and F1-score for the positive class, together with confusion matrix analysis. While Accuracy and ROC-AUC provide a general indication of model discrimination capability, these metrics may be misleading in the presence of class imbalance [24]. Therefore, recall for the positive class was emphasized, as it reflects the model's ability to correctly identify individuals with heart disease and to minimize false negative predictions, which is critical in clinical risk assessment scenarios [23]. Precision and F1-score were additionally reported to capture the trade-off between sensitivity and prediction reliability.

VII. Results

This section presents the experimental results obtained from the applied machine learning models. Model performance is evaluated under unbalanced and balanced learning settings, followed by a comparative analysis of the proposed novelty-based feature grouping strategy.

A. Baseline Model Performance under Unbalanced Learning

Logistic Regression, Random Forest, and XGBoost models were initially trained using the original class distribution of the dataset. As summarized in Table I, all three models achieved relatively high accuracy values, exceeding 0.89. However, these results are largely driven by correct classification of the majority (non-heart disease) class.

Despite strong accuracy and ROC-AUC values, recall for the positive class remained critically low for all unbalanced models. Logistic Regression and XGBoost achieved recall values below 0.06, indicating that the majority of heart disease cases were incorrectly classified as negative. This behavior is a well-known limitation of machine learning models trained on highly imbalanced medical datasets, where accuracy-based evaluation may produce misleading conclusions [10], [24], [25].

Random Forest demonstrated a marginally higher recall compared to the other baseline models; however, its performance remained insufficient for reliable clinical risk detection, consistent with previous observations in cardiovascular prediction studies [8], [26].

TABLE I
Unbalanced model performance comparison

Model	Accuracy	ROC-AUC	Recall_+	Precision_+	F1_+
LogReg_base	0.9143	0.8242	0.0537	0.4949	0.0969
RF_base	0.8973	0.7549	0.1264	0.2791	0.1740
XGB_base	0.9145	0.8252	0.0316	0.5134	0.0595

B. Impact of Class Balancing (Partial Novelty)

To address the limitations observed under unbalanced learning, class-weighted training strategies were applied to all models. The resulting balanced model performances are reported in Table II.

Class balancing led to a substantial improvement in recall for the positive class across all evaluated models. Balanced Logistic Regression and XGBoost achieved recall values of approximately 0.78, representing more than a tenfold increase compared to their unbalanced counterparts. This improvement confirms the effectiveness of imbalance-aware learning approaches in reducing false negative predictions in healthcare applications [10], [23].

Although balanced training resulted in a decrease in overall accuracy, ROC-AUC values remained stable across both learning settings, indicating preserved discriminative capability. Similar trade-offs between accuracy and sensitivity have been widely reported in medical machine learning literature and are considered acceptable when early disease detection is prioritized [8], [25].

TABLE II
Balanced model performance comparison

Model	Accuracy	ROC-AUC	Recall_+	Precision_+	F1_+
LogReg_bal	0.7278	0.8245	0.7770	0.2081	0.3283
RF_bal	0.8894	0.7521	0.1416	0.2460	0.1797
XGB_bal	0.7253	0.8252	0.7845	0.2077	0.3284

C. ROC Curve Analysis

ROC curve analysis further supports the effectiveness of balanced learning strategies. Figures 8 and 9 illustrate the ROC curves of unbalanced and balanced models, respectively.

Balanced Logistic Regression and XGBoost models maintained strong ROC-AUC values while significantly improving sensitivity. These findings demonstrate that class balancing enhances the practical utility of predictive models without compromising their overall classification capability [21], [24].

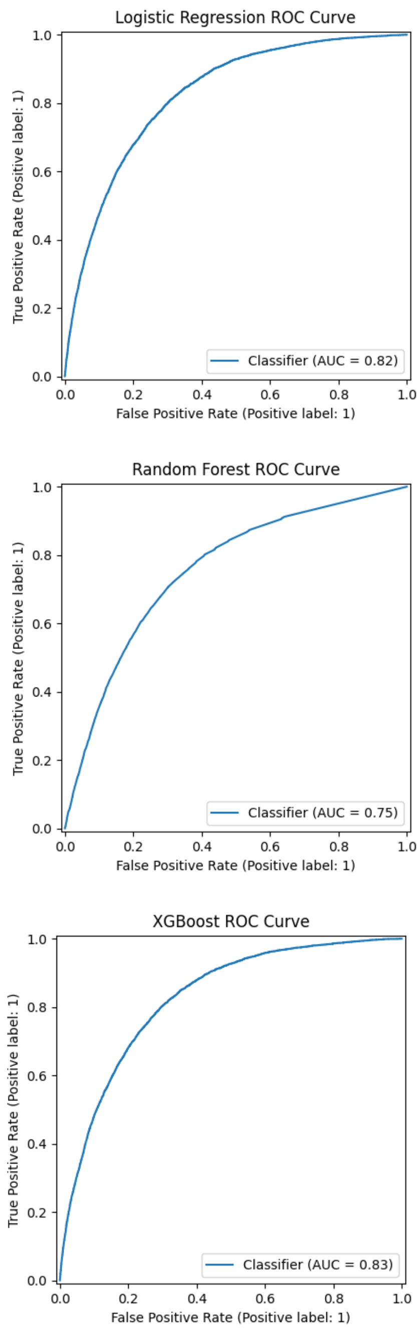

Fig. 8. ROC curves of unbalanced models

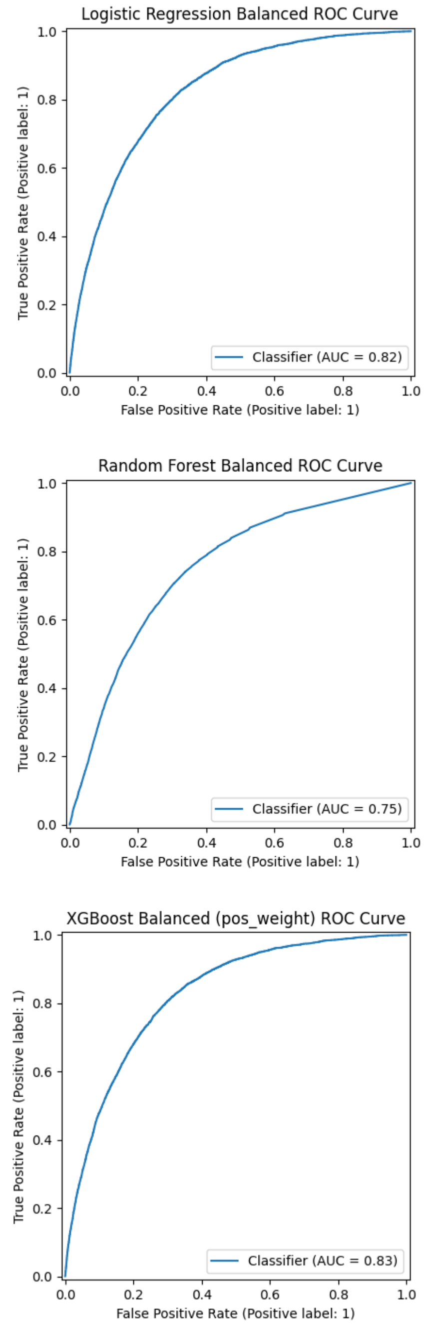

Fig. 9. ROC curves of balanced models

D. Novelty-Based Feature Group Comparison (Novelty)

The primary novelty of this study is the structured comparison of two balanced feature group configurations. Model A incorporates lifestyle, sleep, stress, and demographic features, while Model B extends this configuration by including medical history variables.

Both configurations were evaluated using balanced Logistic Regression and balanced XGBoost models. As shown in Table III, Model B consistently outperformed Model A across most evaluation metrics. The inclusion of medical history features led to improvements in ROC-AUC, precision, and F1-score, while maintaining high recall values.

These results align with existing epidemiological evidence highlighting the strong association between prior clinical conditions and cardiovascular risk [3], [5], [6]. At the same time, the competitive performance of Model A demonstrates that lifestyle and behavioral factors alone can provide meaningful predictive signals, supporting preventive risk modeling approaches based on non-clinical data [11]–[13].

Overall, the results confirm that the proposed novelty-based methodology effectively addresses class imbalance and highlights the complementary roles of lifestyle behaviors and medical history in heart disease prediction.

Table III presents the comparative results.

TABLE III
NOVELTY-BASED MODEL COMPARISON (BALANCED)

Model	Accuracy	ROC-AUC	Recall_+	Precision_+	F1_+
Model A–LogReg_bal	0.7278	0.8245	0.7770	0.2081	0.3283
Model A–XGB_bal	0.7253	0.8252	0.7845	0.2077	0.3284
Model B–LogReg_bal	0.7487	0.8372	0.7750	0.2223	0.3455
Model B–XGB_bal	0.7462	0.8371	0.7812	0.2215	0.3451

The inclusion of medical history features consistently improved ROC-AUC, precision, and F1-score values, while maintaining high recall. These results confirm that integrating clinical background information enhances heart disease risk prediction under balanced learning conditions.

VIII. DISCUSSION

This study investigated heart disease risk prediction using supervised machine learning models under highly imbalanced data conditions. The discussion focuses on three key aspects: (i) the impact of class imbalance on predictive performance, (ii) the effectiveness of imbalance-aware learning strategies, and (iii) the contribution of structured feature grouping, particularly the inclusion of medical history variables.

A. Effect of Class Imbalance on Model Performance

The baseline experiments clearly demonstrate the limitations of training standard machine learning models on imbalanced healthcare datasets. Although Logistic Regression, Random Forest, and XGBoost achieved high accuracy and ROC-AUC values under unbalanced learning, their recall for the positive class was critically low. This indicates that the majority of heart disease cases were misclassified as negative, despite seemingly strong overall performance.

Such behavior is consistent with prior findings showing that accuracy and ROC-AUC alone may provide misleading assessments when class distributions are skewed [10], [24], [25]. In clinical risk prediction, false negatives carry significant consequences, as undetected high-risk individuals may not receive timely intervention. Therefore, the baseline results highlight the necessity of adopting imbalance-aware evaluation and learning strategies in cardiovascular disease modeling.

B. Impact of Balanced Learning Strategies (Partial Novelty)

Applying class-weighted learning substantially improved the sensitivity of all evaluated models. Balanced Logistic Regression and balanced XGBoost achieved recall values close to 0.78, representing a significant improvement over their unbalanced counterparts. This demonstrates that imbalance-aware training effectively mitigates the bias toward the majority class and enhances the detection of clinically relevant cases.

While balanced training led to a reduction in overall accuracy, ROC-AUC values remained largely stable, indicating that the models preserved their discriminative capability. This trade-off between accuracy and recall has been widely documented in medical machine learning studies and is generally considered acceptable when the primary objective is early disease identification [8], [23]. The results confirm that prioritizing recall for the positive class is a more appropriate strategy for cardiovascular risk prediction than maximizing accuracy alone.

C. Model-Specific Observations

Among the evaluated models, Logistic Regression and XGBoost consistently outperformed Random Forest in balanced learning settings. While Random Forest benefited marginally from class weighting, its recall remained comparatively low. This may be attributed to the model's tendency to favor majority-class patterns in highly skewed datasets, particularly when decision thresholds are not explicitly optimized [26].

In contrast, XGBoost demonstrated robust performance across both unbalanced and balanced configurations, aligning with previous studies that highlight its effectiveness in structured healthcare data [8], [21]. Logistic Regression also showed competitive results, reinforcing its suitability for clinical decision support due to its simplicity, stability, and interpretability [27].

D. Contribution of Feature Grouping and Medical History (Novelty)

A key contribution of this work is the structured comparison of two balanced feature group configurations. The results indicate that Model B, which incorporates medical history variables in addition to lifestyle, sleep, stress, and demographic features, consistently outperformed Model A across ROC-AUC, precision, and F1-score metrics.

The inclusion of medical history features such as prior stroke, diabetes, and kidney disease enhanced predictive reliability without compromising recall. This finding is consistent with epidemiological evidence identifying pre-existing conditions as strong determinants of cardiovascular risk [3], [5], [6]. At the same time, the strong performance of Model A demonstrates that non-clinical behavioral and lifestyle factors can provide meaningful predictive signals, supporting the feasibility of early risk assessment using survey-based data [11]–[13].

E. Clinical and Practical Implications

From a clinical perspective, the results emphasize the importance of recall-oriented model evaluation in heart disease prediction. Models optimized solely for accuracy may fail to identify high-risk individuals, whereas imbalance-aware approaches can substantially improve early detection. Furthermore, the proposed feature grouping strategy offers a flexible framework that can be adapted to different data availability scenarios, ranging from population-level screening to clinically enriched datasets.

Overall, the findings suggest that combining imbalance-aware learning with structured feature selection yields more reliable and clinically meaningful prediction models for cardiovascular disease risk assessment.

IX. Conclusion

In this study, heart disease risk prediction was investigated using supervised machine learning models under highly imbalanced data conditions. The experimental results demonstrated that standard learning approaches, although achieving high accuracy, fail to reliably identify individuals with heart disease due to severe class imbalance.

By incorporating imbalance-aware learning strategies, substantial improvements in recall for the positive class were achieved, confirming the importance of sensitivity-oriented evaluation in clinical risk assessment. Furthermore, the proposed novelty-based feature grouping strategy showed that integrating medical history variables significantly enhances predictive reliability while preserving strong discriminative performance.

The findings highlight that lifestyle and behavioral factors alone can provide meaningful early risk signals, while the inclusion of clinical history further strengthens model robustness. Overall, this work emphasizes the necessity of combining imbalance-aware learning with structured feature selection to develop clinically meaningful and practical cardiovascular risk prediction models. These results may support future population-level screening and decision-support systems aimed at early detection and prevention of heart disease [28]–[30].

References

[1] Our World in Data, "Cardiovascular diseases," 2023, available: https://ourworldindata.org/cardiovascular-diseases.

[2] World Heart Federation, "What is cardiovascular disease (cvd)?" 2023, available: https://world-heart-federation.org/what-is-cvd/.

[3] E. J. e. a. Benjamin, "Heart disease and stroke statistics," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[4] A. S. e. a. Go, "Heart disease and stroke statistics," *Circulation*, vol. 127, no. 1, pp. e6–e245, 2013.

[5] R. B. e. a. D'Agostino, "General cardiovascular risk profile for use in primary care," *Circulation*, vol. 117, no. 6, pp. 743–753, 2008.

[6] G. A. e. a. Mensah, "Risk factors and cardiovascular disease," *Heart*, vol. 103, no. 18, pp. 1451–1458, 2017.

[7] L. Wang, R. Fan, C. Zhang, L. Hong, T. Zhang, Y. Chen, K. Liu, Z. Wang, and J. Zhong, "Applying machine learning models to predict medication nonadherence in crohn's disease maintenance therapy," *Patient Preference and Adherence*, vol. 14, pp. 917–926, 2020.

[8] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Machine learning prediction in cardiovascular diseases," *Journal of the American College of Cardiology*, vol. 72, no. 20, pp. 2587–2599, 2018.

[9] Centers for Disease Control and Prevention, "Behavioral risk factor surveillance system," 2023, available: https://www.cdc.gov/brfss.

[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] F. W. e. a. Booth, "Lack of exercise is a major cause of chronic diseases," *Comprehensive Physiology*, vol. 2, no. 2, pp. 1143–1211, 2012.

[12] F. P. e. a. Cappuccio, "Sleep duration and cardiovascular outcomes," *European Heart Journal*, vol. 32, no. 12, pp. 1484–1492, 2011.

[13] A. Steptoe and M. Kivimäki, "Stress and cardiovascular disease," *Nature Reviews Cardiology*, vol. 9, no. 6, pp. 360–370, 2012.

[14] 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.

[15] C. e. a. Krittanawong, "Machine learning prediction in cardiovascular diseases," *Journal of the American College of Cardiology*, vol. 72, no. 20, pp. 2587–2599, 2018.

[16] G. A. e. a. Mensah, "Risk factors and cardiovascular disease," *Heart*, vol. 103, no. 18, pp. 1451–1458, 2017.

[17] F. P. e. a. Cappuccio, "Sleep duration and cardiovascular outcomes," *European Heart Journal*, vol. 32, no. 12, pp. 1484–1492, 2011.

[18] A. Steptoe and M. Kivimäki, "Stress and cardiovascular disease," *Nature Reviews Cardiology*, vol. 9, no. 6, pp. 360–370, 2012.

[19] R. B. e. a. D'Agostino, "General cardiovascular risk profile," *Circulation*, vol. 117, no. 6, pp. 743–753, 2008.

[20] G. A. e. a. Mensah, "Risk factors and cardiovascular disease," *Heart*, vol. 103, no. 18, pp. 1451–1458, 2017.

[21] 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.

[22] 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.

[23] E. W. Steyerberg, *Clinical Prediction Models*, 2nd ed. Springer, 2019.

[24] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[25] T. Saito and M. Rehmsmeier, "The precision–recall plot is more informative than the roc plot on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[28] World Health Organization, "Cardiovascular diseases (cvds)," 2023, available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[29] S. S. e. a. Virani, "Heart disease and stroke statistics—2023 update," *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023.

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

https://github.com/KeremKuru22/The-Effects-of-Sleep-Stress-and-Lifestyle-on-Heart-Disease-Evidence-from-BRFSS-Data