Leveraging Selected Lifestyle Factors from NHANES Data for Chronic Heart Disease Risk Prediction

2025 Intermountain Engineering, Technology and Computing (IETC) | 979-8-3315-1282-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IETC64455.2025.11039383

1st Sharmin Nahar
Dept. of Computer Science
Louisiana State University Shreveport
Shreveport, USA
nahars78@lsus.edu

2nd Dr. Zhonghui (John) Wang
Dept. of Computer Science
Louisiana State University Shreveport
Shreveport, USA
zhonghui.wang@lsus.edu

3rd Dr. Subhajit Chakrabarty
Dept. of Computer Science
Louisiana State University Shreveport
Shreveport, USA
subhajit.chakrabarty@lsus.edu

4th Dr. Xi Jin
Dept. of Kinesiology and Health Science
Louisiana State University Shreveport
Shreveport, USA
xi.jin@lsus.edu

***Abstract*—Chronic Heart Disease (CHD) remains a major global health concern, necessitating early detection strategies for effective intervention. This study employs machine learning (ML) techniques to predict CHD risk using selected lifestyle factors from NHANES data. A comprehensive feature selection process was conducted to enhance predictive accuracy and interpretability. Among various ML models evaluated, XGBoost demonstrated the best performance, achieving an accuracy of 85.94 %, an F1-score of 85.89, and an AUC-ROC of 0.92, underscoring the relevance of lifestyle factors in CHD assessment. These findings highlight the potential of ML-driven methodologies for improving early diagnosis and preventive healthcare strategies.**

***Index Terms*—chronic heart disease, machine learning, lifestyle factors, risk prediction, NHANES**

I. Introduction

Chronic Heart Disease (CHD) is a leading cause of mortality worldwide, contributing to millions of deaths annually. According to the World Health Organization (WHO), cardiovascular diseases, including CHD, account for nearly 17.9 million deaths each year, representing approximately 32% of all global deaths. Early detection and timely intervention are crucial in mitigating CHD risks, as delays in diagnosis and treatment can lead to severe complications such as heart failure, stroke, and sudden cardiac death.

Traditional diagnostic methods primarily rely on clinical assessments, medical history, and physiological measurements such as cholesterol levels, blood pressure, and electrocardiographic (ECG) findings. However, these approaches often require medical expertise, are resource-intensive, and may not be accessible to all populations. More importantly, they tend to overlook the profound impact of lifestyle behaviors on CHD progression.

This study introduces a machine learning-based approach that focuses on non-invasive, lifestyle-driven risk assessment, using accessible self-reported data. The proposed model aims to serve as a potential self-screening tool for early CHD risk detection—particularly in contexts where individuals may lack access to clinical testing. By leveraging behavioral and environmental data from the National Health and Nutrition Examination Survey (NHANES), our goal is to provide a scalable, interpretable, and cost-effective alternative for heart disease awareness and prevention.

Unlike traditional models that prioritize clinical diagnostics, this work centers on lifestyle attributes such as physical activity, sleep patterns, mental health indicators, and substance use. The methodology involves rigorous data preprocessing, feature selection, and classifier evaluation. Through this process, we demonstrate that lifestyle patterns alone can yield strong predictive performance, offering a valuable complementary tool for public health initiatives.

The rest of this paper is organized as follows: Section II reviews existing studies on heart disease prediction using machine learning. Section III describes the dataset and variables used in this study. Section IV details the methodology, including data preprocessing, feature selection, and model evaluation techniques. Section V presents the model performance results. Finally, Section VI discusses the conclusion, limitations, and challenges encountered during the study.

II. Related Work

Machine learning (ML) has become a powerful approach in predictive healthcare, offering data-driven alternatives to

traditional CHD diagnostics, which are often time-consuming and resource-intensive.

A. Machine Learning for Heart Disease Prediction

Numerous studies have applied ML models using clinical and demographic data to predict CHD risk. Patel et al. [12] and Dwivedi [8] found SVM and logistic regression to be among the most accurate classifiers. Bahrami and Shirvani [3] emphasized the role of preprocessing and feature selection.

Ensemble methods have proven particularly effective. Fawagreh et al. [9] highlighted Random Forest's ability to handle high-dimensional data, while Bentejac et al. [4] and Dorogush et al. [7] demonstrated the advantages of boosting techniques like XGBoost and CatBoost for managing complex feature interactions. Devale and Mujawar [14] proposed a hybrid K-Means and Naïve Bayes model to reduce noise and improve classification. XGBoost, introduced by Chen and Guestrin [5], remains a widely adopted framework in healthcare for its speed, robustness, and accuracy.

B. The Role of Lifestyle Factors in CHD Prediction

Beyond clinical indicators, lifestyle factors are crucial for CHD prediction. Indrakumari et al. [11] and Deepika and Seema [6] analyzed behaviors like smoking, diet, and physical activity. Sharmila and Chellammal [13], and Shah et al. [15], demonstrated that obesity, stress, and inactivity significantly influence CHD risk. Foundational work from the 1970s and 1980s [19], [20] on psychosocial contributors has been validated by modern ML models.

C. Feature Selection, Data Imbalance, and Optimization Techniques

Robust feature selection is key to effective CHD prediction. Alizadehsani et al. [2] and Ge et al. [10] leveraged techniques like the Spearman Rank Correlation [1] and RFE to enhance model reliability. Data imbalance, a common issue, has been addressed through undersampling [16] and oversampling techniques like SMOTE. For tuning models, Bayesian Optimization has been effective in improving performance, as demonstrated by Taneja et al. [17] and Rahhal et al. [18].

D. Contributions of This Study

While prior research has demonstrated the efficacy of ML in CHD prediction, most studies have predominantly relied on clinical diagnostic features. This study builds upon existing work by shifting the focus toward lifestyle factors, leveraging a public dataset to assess behavioral attributes in CHD risk prediction. By integrating feature selection techniques, handling class imbalance through undersampling, and optimizing model performance, this study aims to provide a more interpretable and practical approach to CHD risk assessment.

III. DATASET

The dataset used in this study was sourced from the open-access National Health and Nutrition Examination Survey (NHANES), publicly available at: https://wwwn.cdc.gov/nchs/nhanes/default.aspx. We focused on survey cycles from 2005 to 2020, excluding the pandemic years to maintain data consistency. The COVID-19 period introduced disruptions in lifestyle, healthcare access, and disease reporting, which could skew predictive modeling. Therefore, we retained only pre-pandemic data to enhance model reliability.

The selected dataset includes a wide range of variables representing demographic information, lifestyle habits, medical history, and mental health indicators. These include: Age, Sex, Marital Status, Healthcare Access, BMI, Calories, Sodium, Saturated Fat, Added Sugar, Fiber, Caffeine, Junk Food Consumption, Vigorous Work, Moderate Work, Active Transport (e.g., walking/biking), Vigorous Recreation, Moderate Recreation, Minutes of Sedentary Activity, Depression, Stress, Alcohol Consumption, Smoking Status, Smoke Exposure, Sleep Duration, Sleep Disorders, Employment Status, Blood Lead Level, Diabetes Status, and derived Hypertension status.

Hypertension was determined based on either a systolic blood pressure reading $\geq$140 mmHg (BPXOSY1) or self-reported physician diagnosis (BPQ020 = 1). These features were chosen for their relevance to chronic heart disease (CHD) and their accessibility through self-reporting, aligning with the study's goal of developing a non-invasive, lifestyle-based risk prediction model.

The target variable, CHD Status, was constructed using self-reported responses from the NHANES Medical Conditions Questionnaire (MCQ). A participant was labeled as CHD-positive (CHD Status = 1) if they reported having any of the following physician-diagnosed conditions: congestive heart failure (MCQ160E = 1), coronary heart disease (MCQ160C = 1), or angina pectoris (MCQ160D = 1). Otherwise, they were labeled as CHD-negative (CHD Status = 0). This definition provides a clinically meaningful basis for binary classification in the predictive modeling process.

The NHANES Medical Conditions (MCQ) questionnaire module serves as the source of this data, providing self-reported information on physician-diagnosed health conditions. Since these conditions are strong indicators of cardiovascular disease risk, defining CHD Status in this manner ensures a clear and clinically relevant classification for predictive modeling.

IV. METHODS

A. Data Preprocessing and EDA (Exploratory Data Analysis)

The obtained files from the NHANES database were originally in XPT (SAS Transport Format) and were converted into Pandas DataFrames using Python for further processing.

Before applying machine learning models, extensive preprocessing was conducted to ensure data quality and model reliability. The raw NHANES dataset consisted of 76,496 rows and 34 columns, which, after exploratory data analysis (EDA) and preprocessing, was reduced to 7,004 rows and 27 columns while preserving meaningful information. The preprocessing steps are detailed below.

Data Cleaning and Validity Checks: Initial dataset exploration was performed to check for missing values, data types, and statistical distributions. Non-logical entries, such as age values of 0 or below 16, were removed, as NHANES primarily measures different variables for children and adults separately, and CHD is rare in younger populations.The dataset includes individuals aged 16 to 85 years after filtering out non-logical entries (ages 0-15). Additionally, NHANES does not provide exact ages for individuals above 85 years; instead, they are grouped into a single ”85+” category. This limits the granularity of age-related analyses, as specific risk variations among elderly participants cannot be examined in detail.

Encoding Categorical Variables: The Sex column was encoded into numerical values (e.g., Male $\rightarrow$ 0, Female $\rightarrow$ 1) to facilitate model training.

Feature Selection: To ensure that the model learned from lifestyle factors rather than direct clinical diagnoses, we identified key risk factors relevant to heart disease.

Handling Missing Values: Columns with more than 60 % missing values were analyzed, and Smoke Exposure was removed due to excessive missing data. For other missing values, an advanced imputation technique KNN Imputer was used. Data was processed in batches, imputed, and then concatenated back to maintain consistency.

Outlier Detection and Removal: Outliers in continuous numerical variables were identified using boxplots and handled using the Interquartile Range (IQR) method to prevent extreme values from biasing model training.

Addressing Class Imbalance: The dataset exhibited significant class imbalance, with 6.5% CHD cases and 93.5% non-CHD cases as shown in Fig. 1. Undersampling was applied instead of oversampling to preserve data integrity, ensuring that all data remained authentic without introducing synthetic examples.

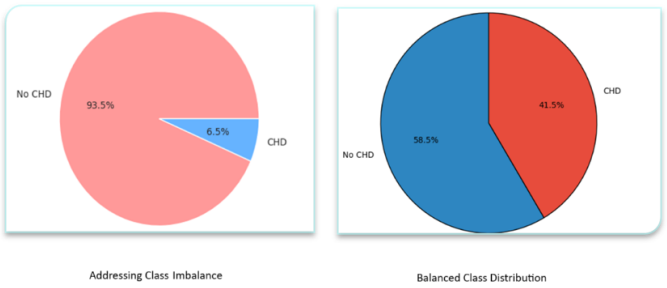

Fig. 1. Class Balancing

Correlation and Multicollinearity Handling: A Pearson correlation matrix was computed to detect multicollinearity among features. The following highly correlated variables were identified and processed: Sodium and Calories, Saturated Fat and Calories. To reduce redundancy, Sodium and Saturated Fat were dropped while retaining Calories, which had a stronger correlation with the target variable.

B. Machine Learning Models

To effectively predict chronic heart disease and identify influential lifestyle factors, we employed a diverse set of machine learning models selected based on their interpretability, ability to handle non-linear relationships, and relevance to the nature of the dataset.

- **Logistic Regression (LR)** was chosen for its simplicity and interpretability, providing clear insight into how individual predictors influence disease risk through model coefficients.
- **K-Nearest Neighbors (KNN)** was included to explore whether similar lifestyle profiles correspond to similar risk levels. However, we acknowledge its limitations, such as sensitivity to data distribution and lack of model-based learning, and did not rely on it as a primary estimator.
- **Support Vector Machine (SVM)** was applied to capture complex, high-dimensional relationships in the data, useful for detecting subtle patterns not captured by linear models.
- **Decision Tree (DT**) offered a transparent, rule-based structure for evaluating the hierarchical significance of features, aligning well with clinical decision-making frameworks.
- **Random Forest (RF)** improved robustness by aggregating multiple decision trees, reducing overfitting, and offering reliable feature importance scores across various data subsets.
- **XGBoost (XGB)** was selected for its high accuracy and refined feature selection through gradient boosting, helping uncover less obvious but influential factors in heart disease risk.

This model selection strategy allowed us to cross-validate findings, balance interpretability with predictive power, and ensure that key lifestyle-related predictors were consistently highlighted across methods.

C. Model Training Procedure

Following the selection of machine learning models, a structured procedure was implemented to train, evaluate, and optimize the models for CHD prediction. The key steps in the modeling process are outlined below:

Data Loading and Target Definition: The preprocessed dataset was loaded, and the target variable (CHD Status) was separated from the feature set.

Train-Test Splitting: The dataset was split into 80% training and 20 % testing subsets to ensure a fair evaluation of model performance.

Feature Scaling: Standardization was applied using StandardScaler, transforming numerical features to have zero mean and unit variance to improve model stability and convergence.

Model Training: Multiple machine learning classifiers were trained to compare their performance in predicting CHD risk.

Hyperparameter Tuning: GridSearchCV was employed for hyperparameter optimization, ensuring that each model was fine-tuned for optimal performance.

Prediction and Evaluation: The trained models were used to make predictions on the test dataset. Performance was assessed using the following evaluation metrics: Accuracy, F1-Score, Precision, Recall, AUC-ROC Score, Confusion Matrix.

Further Model Optimization: The best-performing model was further optimized to maximize predictive performance, ensuring that it captured the most relevant CHD risk factors effectively.

V. RESULTS OF MODEL PERFORMANCE

The performance of multiple machine learning models was evaluated to identify the most effective classifier for CHD prediction. The evaluation metrics included accuracy, F1-score, precision, recall, and AUC-ROC score, providing a comprehensive comparison.

A. Best Performing Model: Among the tested models, XGBoost achieved the highest accuracy (84.23) and F1-score (84.20), making it the most effective model for predicting CHD occurrence.

B. Precision and Recall Analysis: XGBoost demonstrated superior performance in precision and recall for both classes (0 and 1), indicating its effectiveness in minimizing false positives while correctly identifying true positives. Decision Tree and Random Forest models also performed well but were slightly behind XGBoost in terms of predictive accuracy and generalization.

C. Model Recommendation: Due to its high accuracy, F1-score, and balanced precision-recall tradeoff, XGBoost is the recommended model for CHD prediction. For applications requiring simpler implementation with reasonable accuracy and interpretability, Decision Tree or Random Forest can be considered as viable alternatives.

D. Performance Comparison Table: The table below presents a comparative analysis of different machine learning models based on key evaluation metrics:

Model Comparison Table for CHD Prediction

Model	Accuracy	F1-Score	Precision (Class 0)	Recall (Class 0)	Precision (Class 1)	Recall (Class 1)
Logistic Regression	0.7709	0.7717	0.82	0.78	0.72	0.76
KNN	0.7580	0.7593	0.82	0.75	0.69	0.77
SVM	0.7709	0.7721	0.83	0.76	0.71	0.78
Decision Tree	0.7944	0.7927	0.80	0.86	0.78	0.71
Random Forest	0.7901	0.7908	0.83	0.80	0.74	0.78
XGBoost	**0.8423**	**0.8420**	**0.86**	**0.87**	**0.82**	**0.80**

Fig. 2. Model Comparison Table

This performance evaluation highlights XGBoost as the most suitable model for CHD risk assessment, offering a robust balance of accuracy and interpretability. Further improvements can be explored through advanced feature selection and model optimization techniques.

E: Model Optimization: To enhance the predictive performance of the XGBoost classifier, we implemented a comprehensive optimization strategy combining feature selection and hyperparameter tuning. Feature selection was carried out using Recursive Feature Elimination (RFE), which iteratively removed less informative variables and retained the top 12 most impactful features. Subsequently, Bayesian Optimization was employed to fine-tune key hyperparameters such as learning rate, maximum depth, and the number of estimators. This two-step optimization process significantly improved model performance, achieving an accuracy of 85.94 and an F1-score of 85.89, while also increasing model efficiency and interpretability. As shown in Fig. 3 the results confirmed that the optimized XGBoost model outperformed previous iterations, achieving a better balance between precision and recall while maintaining robust predictive capability.

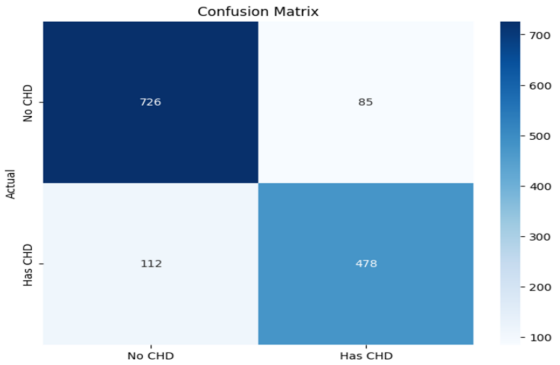

Fig. 3. Confusion Matrix for XGBoost

ROC AUC curve: The AUC-ROC curve is an essential tool used for evaluating the performance of binary classification models. It plots the True Positive VS the False Positive showing how well a model can distinguish between two classes such as positive and negative outcomes. It provides a graphical representation of the model's ability to distinguish between two classes like positive class for presence of a disease and negative class for absence of a disease. The ROC curve and AUC score of 0.92 confirm that the machine learning model is highly effective at predicting CHD based on selected lifestyle factors.

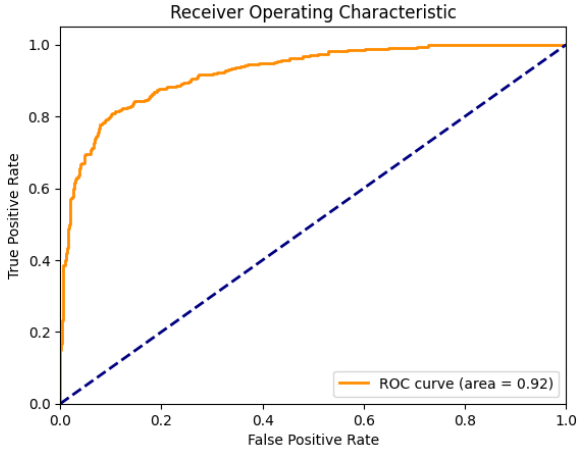

Fig. 4. AUC-ROC curve

Findings from the ROC-AUC Curve:

Strong Model Performance: The ROC-AUC score is 0.92, indicating that the model has a high discriminatory power in distinguishing between CHD-positive and CHD-negative cases.

Excellent True Positive Rate: As shown in Fig. 4 the curve shows a steep rise at the beginning, meaning the model correctly identifies a high proportion of true positives (high sensitivity/recall) while keeping false positives relatively low.

Minimal Overfitting: A smooth and consistently high curve without abrupt fluctuations suggests that the model is well-optimized and generalizes well to unseen data.

Better Than Random Guessing: The diagonal dashed line (AUC = 0.5) represents random classification. Since the ROC curve is well above this line, the model is significantly better than random guessing.

F: Feature Importance:

To enhance model interpretability and efficiency, Recursive Feature Elimination (RFE) with XGBoost was utilized to identify the most influential features contributing to chronic heart disease (CHD) prediction. The RFE process iteratively removed less relevant features and selected an optimal subset that maximized model performance.

(a) Optimal Feature Subset Selection: The best-performing model was achieved using 12 selected features, as determined by RFE. The model using this optimal subset yielded: **Accuracy: 85.94, F1-Score: 85.89**. These metrics indicate that the refined model maintained high predictive accuracy while ensuring a balance between precision and recall.

(b) Top 12 Features Affecting CHD: the final 12 most significant features (shown in Fig. 5) influencing CHD prediction are:

Hypertension, **Age**, **Diabetes**, **Employment Status**, **Minutes of Sedentary Activity**, **Smoking Status**, **Alcohol Consumption**, **Sleep Disorders**, **Sex**, **Active Transport (Walking/Biking)**, **Depression**, **Sleep Duration**

1. **Hypertension** (High Blood Pressure) – High blood pressure increases heart workload, leading to vascular damage and CHD risk [6]. 2. **Age** – Arterial stiffening and cumulative risk factors with age elevate CHD susceptibility [6], [15]. 3. **Diabetes** – Elevated blood glucose damages blood vessels and promotes atherosclerosis [6]. 4. **Employment Status** – Job instability and low socioeconomic status increase stress and hinder healthy lifestyles [13]. 5. **Minutes Sedentary Activity** – Physical inactivity correlates with obesity, poor circulation, and higher CHD incidence [11]. 6. **Smoking Status** – Smoking damages vascular linings, raises blood pressure, and accelerates heart damage [11], [15]. 7. **Alcohol Consumption** – Excessive intake raises blood pressure and cholesterol, elevating CHD risk [11]. 8. **Sleep Disorders** – Linked to elevated stress hormones and disrupted cardiovascular regulation [6]. 9. **Sex (Gender)** – Males typically exhibit higher early-age CHD risk, while postmenopausal women face increased vulnerability [15]. 10. **Active Transport** (Walking, Biking, etc.) – Walking/biking enhances cardiovascular health and reduces sedentary effects [11]. 11. **Depression** – Associated with unhealthy behaviors and physiological stress responses detrimental to heart health [6], [15]. 12. **Sleep Duration** – Abnormal sleep (short or long) is linked to inflammation, metabolic dysfunction, and cardiovascular strain [6].

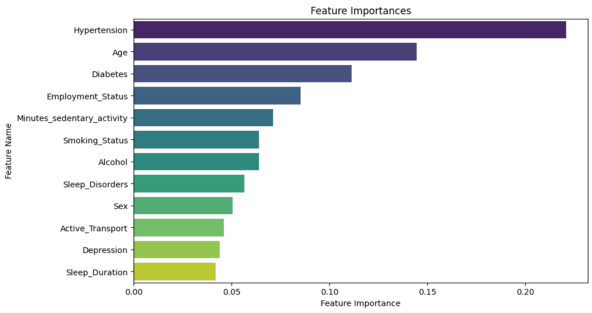

Fig. 5. Final Selected Lifestyle Features (12) for CHD risk prediction

Our model findings align with well-established medical literature, highlighting that both biological factors (hypertension, diabetes, age) and lifestyle choices (sedentary behavior, smoking, alcohol, sleep) play a critical role in predicting CHD risk.

(c) Performance Insights: The optimized feature set shows a strong link between lifestyle factors and CHD risk. It enhanced both interpretability and model efficiency, supporting its value in machine learning-based CHD prediction.

VI. CONCLUSION

This study successfully developed an optimized machine learning model for chronic heart disease (CHD) prediction using lifestyle factors from the NHANES dataset. The Recursive Feature Elimination (RFE) with XGBoost method identified 12 key features that significantly influence CHD risk, including hypertension, diabetes, sedentary activity, and sleep disorders. The final model achieved an accuracy of 85.94 and an F1-score of 85.89, demonstrating its strong predictive capability.

Challenges and Mitigations: Throughout the research, several challenges were encountered.

- Data Issues: The dataset contained missing values and an imbalanced target class distribution.
- Feature Engineering: Highly correlated features required careful selection to improve model performance.
- Model Optimization: Finding the best hyperparameters was crucial to maximizing accuracy and stability.

These challenges were addressed through data preprocessing, feature selection, and Bayesian Optimization for hyperparameter tuning, ensuring a robust predictive model.

Ethical and Practical Considerations: While the proposed lifestyle-based model shows promise for early CHD risk prediction, deploying it in real-world applications raises several ethical and practical concerns that must be addressed.

Comparison with Clinical Methods: Although clinical screening tools are often highly accurate, they typically require access to healthcare, lab tests, and medical professionals. The proposed model offers a more accessible and cost-effective

alternative, especially for at-risk individuals in underserved or remote areas. However, it is not intended to replace clinical assessments but rather to serve as a preliminary self-assessment tool.

False Positives: In the event of a false positive prediction, users may experience unnecessary anxiety. To mitigate this, any deployed version of the model should clearly communicate that results are indicative—not diagnostic—and recommend consulting a healthcare provider for confirmation.

False Negatives: False negatives pose a more significant risk, potentially leading users to ignore real symptoms. To address this, it is essential to include strong disclaimers and educational prompts that advise users not to rely solely on the tool for medical decisions.

Responsible Use: If deployed as a public-facing application or integrated into wellness platforms, the model should include mechanisms such as informed consent, privacy protection for user data, and referral links to clinical resources. Continuous updates and validation with diverse, real-world datasets would also be crucial to ensure fairness and generalizability.

Future Work: While the developed model performs well, there are opportunities for further improvement like

- Testing with more real-world datasets to enhance generalizability,
- Exploring deep learning models for potential accuracy improvements.
- Deploying the model in real-time healthcare applications for early detection and intervention.

Final Remarks: This study highlights the critical role of lifestyle factors in CHD prediction and demonstrates the effectiveness of machine learning in early risk assessment. By refining predictive models and integrating them into healthcare systems, we can advance preventive strategies and improve cardiovascular health outcomes at both individual and population levels.

REFERENCES

[1] Spearman Rank Correlation Coefficient, pages 502–505. Springer New York, New York, NY, 2008.

[2] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi, and N. Sarrafzadegan. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Scientific Data, 6(1):227, 2019.

[3] Boshra Bahrami and Mirsaeid Hosseini Shirvani. Prediction and diagnosis of heart disease by data mining techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2(2):164–168, 2015.

[4] Candice Bentejac, Anna Csörgo, and Gonzalo Martínez-Munoz. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3):1937–1967, 2021.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.

[6] Kumari Deepika and S Seema. Predictive analytics to prevent and control chronic diseases. In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pages 381–386. IEEE, 2016.

[7] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: Gradient boosting with categorical features support. CoRR, abs/1810.11363, 2018.

[8] Ashok Kumar Dwivedi. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10):685–693, 2018.

[9] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: From early developments to recent advancements. Systems Science and Control Engineering, 2(1):602–609, 2014.

[10] Fuchao Ge, Yuntao Ju, Zhinan Qi, and Yi Lin. Parameter estimation of a Gaussian mixture model for wind power forecast error by Riemann L-BFGS optimization. IEEE Access, 6:38892–38899, 2018.

[11] R. Indrakumari, T. Poongodi, and Soumya Ranjan Jena. Heart disease prediction using exploratory data analysis. Procedia Computer Science, 173:130–139, 2020. International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020.

[12] Jaymin Patel, Dr Tejal Upadhyay, and Samir Patel. Heart disease prediction using machine learning and data mining techniques. Heart Disease, 7(1):129–137, 2015.

[13] S. Chellammal and R. Sharmila. A conceptual method to enhance the prediction of heart diseases using data techniques. International Journal of Computer Science and Engineering, May 2018.

[14] P.R. Devale and Sairabi H. Mujawar. Prediction of heart disease using modified K-means and by using Naïve Bayes. International Journal of Innovative Research in Computer and Communication Engineering, 3:10265–10273, October 2015.

[15] Devansh Shah, Samir Patel, and Santosh Kumar Bharti. Heart disease prediction using machine learning techniques. SN Computer Science, 1(6):345, 2020.

[16] Sathees Kumar B and Sharan Monica L. Analysis of cardiovascular disease prediction using data mining techniques. International Journal of Modern Computer Science, 4:55–58, February 2016.

[17] Abhishek Taneja et al. Heart disease prediction system using data mining techniques. Oriental Journal of Computer Science and Technology, 6(4):457–466, 2013.

[18] M.M. Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and R.R. Yager. Deep learning approach for active classification of electrocardiogram signals. Information Sciences, 345:340–354, 2016.

[19] S. Booth-Kewley and H. S. Friedman, "Psychological predictors of heart disease: A quantitative review," Psychological Bulletin, vol. 101, no. 3, pp. 343–362, 1987.

[20] S. M. Sales and J. S. House, "Job dissatisfaction as a possible risk factor in coronary heart disease," J. Chronic Dis., vol. 23, no. 12, pp. 861–873, 1971.