Synthetic Data Generation and Machine Learning for Enhanced Heart Disease Risk Prediction

G. Srija[1], E. Bhargavi[2], G. Sai Teja[3], B. Kotesh[4], P. Anusha[5], B. Santhosh Kumar[6]
[1,2,3,4]UG Scholar, [5]Professor
Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
Email id : gsrijanaidu@gmail.com[1], ebhargavi04@gmail.com[2], saitejagannu@gmail.com[3], rekotesh@gmail.com[4], palagatianushareddy@gmail.com[5], bsanthosh.csegnit@gniindia.org[6]

Article Received: 10 May 2025 *Article Accepted: 28 May 2025* *Article Published: 2 July 2025*

Citation
G. Srija, E. Bhargavi, G. Sai Teja, B. Kotesh, P. Anusha, B. Santhosh Kumar, "Synthetic Data Generation and Machine Learning for Enhanced Heart Disease Risk Prediction", Journal of Next Generation Technology (ISSN: 2583-021X), 5(5), pp. 22-30. July 2025.
DOI: 10.5281/zenodo.16692126

Abstract

According to the World Health Organization (WHO), heart diseases cause 12 million deaths annually, with cardiovascular diseases accounting for half of all deaths in many countries. Early diagnosis of cardiovascular diseases can facilitate lifestyle changes in high-risk patients, thereby reducing complications. This paper employs machine learning techniques for heart disease detection and applies sampling methods to handle unbalanced datasets. The study uses the publicly available Framingham Heart Disease dataset from Kaggle, containing 15 features from 4,239 patient records, to predict the 10-year risk of coronary heart disease (CHD). By applying machine learning models alongside sampling techniques, an accuracy of 99% was achieved in heart disease detection.

Keywords: Heart disease, CHD, Machine learning, sampling.

I. Introduction

A. The Global Burden of Heart Disease

Cardiovascular diseases (CVDs), particularly coronary heart disease (CHD), remain the leading cause of mortality worldwide. According to the World Health Organization (WHO), an estimated 17.9 million deaths annually are attributed to CVDs, accounting for 32% of all global deaths. Among these, CHD is the most prevalent, responsible for nearly 7.4 million deaths per year. In the United States alone, 610,000 people die from heart disease annually, with someone suffering a heart attack every 40 seconds. The economic burden is equally staggering, with CVD-related healthcare costs exceeding $363 billion in the U.S.

The situation is equally alarming in developing nations. For instance, in India, CVDs contribute to 28% of all deaths, with mortality rates rising due to urbanization, sedentary lifestyles, and poor dietary habits. The increasing prevalence of risk factors such as hypertension, diabetes, obesity, and smoking further exacerbates the crisis. Despite advancements in medical technology, early detection remains a challenge, leading to preventable complications and fatalities[1]-[3].

B. Why Early Detection Matters

Early diagnosis of heart disease can significantly reduce mortality rates by enabling timely interventions such as:

- Lifestyle modifications (diet, exercise, smoking cessation).
- Pharmacological treatments (statins, antihypertensives).
- Surgical procedures (stents, bypass surgery).

However, traditional diagnostic methods, such as the Framingham Risk Score, rely on statistical models that often lack precision in individualized risk assessment. These models may underestimate or overestimate risk in certain populations, particularly in ethnically diverse cohorts. This limitation underscores the need for more accurate, scalable, and automated prediction tools.

C. The Role of Machine Learning in Heart Disease Prediction

Machine learning (ML), a subset of artificial intelligence (AI), has emerged as a transformative tool in healthcare due to its ability to:

- Analyze large datasets (electronic health records, genomic data, wearable sensor data).
- Identify complex patterns that human clinicians may miss.
- Continuously improve prediction accuracy through iterative learning.

D. How ML Outperforms Traditional Methods

Traditional risk prediction models (e.g., ACC/AHA Pooled Cohort Equations) use linear regression-based approaches, which may fail to capture non-linear relationships between risk factors. In contrast, ML algorithms such as:

- Random Forest (ensemble learning),
- Support Vector Machines (SVM) (kernel-based classification),
- Neural Networks (deep learning),

can model intricate interactions between variables, leading to higher predictive accuracy. For example:

- A study by Weng et al. (2017) showed that ML models improved CVD risk prediction accuracy by 7.6% over traditional methods.
- Alaa et al. (2019) demonstrated that automated ML achieved 77% accuracy in predicting CVD risk using UK Biobank data.

E. Challenges in ML-Based Heart Disease Prediction

Despite its potential, ML faces several challenges in clinical deployment:

1. Imbalanced Datasets: Most heart disease datasets have far fewer high-risk cases than low-risk ones, leading to biased models.
    - Example: The Framingham dataset used in this study has only 15% high-risk patients.
2. Feature Selection: Not all clinical variables are equally predictive. Irrelevant features can decrease model performance.
3. Interpretability: Many advanced ML models (e.g., deep learning) act as "black boxes," making it difficult for clinicians to trust their predictions.

To address these challenges, this study employs sampling techniques (oversampling, SMOTE, ADASYN) and feature engineering to enhance model robustness.

F. The Framingham Heart Study Dataset

This study leverages the Framingham Heart Study dataset, a longitudinal cohort study initiated in 1948, which has become a gold standard in cardiovascular research. The dataset includes:

- 4,239 patient records with 15 clinical and demographic features.
- Target variable: 10-year risk of CHD (binary classification: 1 = high risk, 0 = low risk).

G. Key Features Used in Prediction

Category	**Features**	**Clinical Relevance**
Demographic	Age, Sex, Education	Age and sex are major non-modifiable risk factors.
Behavioral	Smoking status, Alcohol use	Smoking increases CHD risk by **2–4x**.
Clinical	Blood pressure, Cholesterol, BMI	Hypertension and obesity are key modifiable risks.
Medical History	Diabetes, Prior stroke, Medication use	Diabetes doubles CHD risk.

H. Data Imbalance: A Critical Issue

The dataset is highly imbalanced, with only 644 high-risk cases (15.2%) compared to 3,595 low-risk cases (84.8%). Training an ML model on such data without correction would lead to:

- High false-negative rates (missing high-risk patients).
- Over-optimistic accuracy (e.g., a model predicting "low risk" for everyone would still be 85% accurate).

To mitigate this, we apply three sampling techniques:

1. Random Oversampling: Duplicates minority-class samples.
2. SMOTE (Synthetic Minority Oversampling): Generates synthetic high-risk cases.
3. ADASYN (Adaptive Synthetic Sampling): Focuses on "hard-to-classify" borderline cases.

The study's objectives are:

1. To evaluate the performance of ML models in predicting 10-year CHD risk.
2. To compare the effectiveness of Random Oversampling, SMOTE, and ADASYN in handling dataset imbalance.
3. To identify the best-performing algorithm for clinical deployment.

By addressing these objectives, this research aims to provide a data-driven framework for early CHD detection, potentially reducing healthcare costs and improving patient outcomes.

II. Review Of Literature

Jabbar et al. (2013) pioneered the use of genetic algorithms for feature selection in cardiac risk prediction models. Their work with the UCI Heart Disease dataset demonstrated that intelligent feature selection could improve K-NN accuracy by 12 percentage points, identifying age, maximum heart rate, and exercise-induced angina as the most clinically relevant predictors. However, the study's limited sample size (n=303) raised questions about generalizability to broader populations, highlighting the need for validation across larger cohorts.

Sultana et al. (2016) developed a comprehensive five-step data cleaning framework that became foundational for subsequent research. Their systematic approach to handling missing data through kNN imputation and outlier detection using DBSCAN clustering significantly reduced prediction errors by 15% across various algorithms. This work emphasized that model performance depends as much on data quality as on algorithmic sophistication, establishing preprocessing as a critical phase in the machine learning pipeline.

The landmark study by Weng et al. (2017) provided the first large-scale validation of machine learning against traditional risk scores, analyzing 378,256 UK primary care records. Their finding that gradient boosting could outperform conventional methods by 7.6% in AUC metrics came with an important caveat - models required at least 50,000 samples to achieve stable performance. This revelation helped set realistic expectations about data requirements for clinical implementations.

Alaa et al. (2019) explored automated machine learning (AutoML) using Google's framework on the UK Biobank dataset. While achieving 77% accuracy, their models showed poor recall (58%) for high-risk patients, vividly illustrating the class imbalance problem. This study served as a cautionary tale about the limitations of purely automated approaches without specialized handling of imbalanced medical data.

Nakanishi et al. (2018) broke new ground by combining EHR data with coronary calcium scoring from CT imaging. Their neural network architecture, which incorporated attention mechanisms to weight different data modalities, achieved an impressive AUC of 0.89. However, the reliance on expensive imaging limited practical applicability, prompting questions about cost-effective alternatives for widespread screening.

Chen et al. (2020) leveraged data from the Apple Heart Study to demonstrate the potential of consumer wearables. By processing 1.2 billion heart rate measurements from 419,093 participants, their LSTM-based model could detect atrial fibrillation with 71% sensitivity. This study opened new possibilities for continuous, real-time risk monitoring outside clinical settings.

Bunkhumpornpat et al. (2019) conducted the most comprehensive comparison of oversampling methods, evaluating seven SMOTE variants on Framingham data. Their work established ADASYN as superior for preserving data cluster structures while identifying Borderline-SMOTE as the most computationally efficient option. These findings helped guide best practices for handling imbalanced cardiac datasets.

Lundberg et al. (2020) advanced model interpretability through SHAP analysis, quantifying feature importance across multiple studies. Their visualization tools revealed unexpected interaction effects, such as between smoking status and vitamin D levels, providing clinicians with actionable insights beyond simple risk scores.

Caruana et al. (2022) addressed the crucial gap between model accuracy and clinical adoption. Their interpretable generalized additive models achieved 92% approval from practicing cardiologists, demonstrating that explainability could be maintained without significant sacrifice in predictive performance (84% accuracy).

Li et al. (2023) pioneered federated learning for cardiac risk prediction across 23 international hospitals. Their privacy-preserving approach using homomorphic encryption maintained 83% accuracy without requiring data sharing, offering a solution to one of healthcare AI's most persistent challenges - data siloing.

Wang et al. (2023) pushed the boundaries of deployment by implementing quantized neural networks on consumer smartwatches. Achieving real-time predictions with <50ms latency and 0.3mJ energy consumption per inference, this work demonstrated the feasibility of ubiquitous cardiac risk monitoring through wearable devices.

III. Analysis And Conclusion

Table 1: Dataset Overview

Feature	Description	Significance
Source	Framingham Heart Study (Kaggle)	Publicly available benchmark dataset
Records	4,239 patients	Moderate sample size
Features	15 (Age, sex, BP, cholesterol, smoking, etc.)	Covers demographic, clinical, and behavioral factors
Target Variable	TenYearCHD (Binary: 1 = High risk, 0 = Low risk)	Predicts 10-year coronary heart disease risk
Class Distribution	644 high-risk (15.2%), 3,595 low-risk (84.8%)	Highly imbalanced dataset

Table 2: Sampling Techniques Comparison

Technique	Methodology	Advantages	Limitations
Random Oversampling	Duplicates minority class samples randomly	Simple implementation	Risk of overfitting
SMOTE	Generates synthetic samples using k-nearest neighbors	Reduces overfitting	May create noisy samples
ADASYN	Focuses on "hard-to-classify" minority samples	Improves minority class recall	Computationally intensive

Table 3: Model Performance by Sampling Technique

Algorithm	Random Oversampling (Accuracy)	SMOTE (Accuracy)	ADASYN (Accuracy)
Support Vector Machine (SVM)	99%	86%	82.30%
Random Forest	97%	91.30%	90.30%
Extra Tree Classifier	97.80%	91%	90.30%

Gradient Boosting	94.60%	87.80%	87.40%
Decision Tree	91.50%	82%	81.80%
K-NN	79.40%	82%	80.50%
Logistic Regression	67.50%	68.80%	65.70%

Table 4: Precision-Recall Tradeoff

Algorithm	Precision (Avg.)	Recall (Avg.)	Implications
SVM	99.70%	100%	High recall but may lack generalizability due to oversampling artifacts
Random Forest	93%	91.70%	Balanced precision-recall, suitable for clinical use
Naïve Bayes	73%	33.70%	Poor recall indicates failure to capture high-risk cases

Table 5: Comparative Analysis with Prior Studies

Study	Dataset	Best Model	Accuracy	Sampling Used?	Key Difference from Current Study
Alaa et al. (2019)	UK Biobank (423K)	AutoML	77%	No	Larger dataset but lower accuracy
Weng et al. (2017)	UK CPRD (378K)	Gradient Boosting	78.1% (AUC)	No	Focused on AUC, not class imbalance
Mohan et al. (2019)	Cleveland Clinic	Hybrid Random Forest	88.70%	Yes (SMOTE)	Used feature selection but smaller dataset
Current Study	Framingham (4,239)	SVM + Oversampling	99%	Yes (3 techniques)	Achieved highest accuracy via sampling + model tuning

Observations in the study

- The study analyzed machine learning techniques for predicting heart disease risk using the Framingham dataset, which contained 4,239 patient records with significant class imbalance - only 15.2% were high-risk cases.
- Three sampling methods were tested to address this imbalance: random oversampling, SMOTE, and ADASYN. The results showed that random oversampling combined with Support Vector Machines achieved remarkable performance with 99% accuracy

and 100% recall, making it highly effective at identifying high-risk patients. However, this approach may be prone to overfitting due to the artificial duplication of minority class samples.

- Random Forest demonstrated more consistent performance across all sampling techniques, maintaining 90-97% accuracy while balancing precision and recall, suggesting it may be more reliable for real-world clinical applications where generalizability is crucial.

- Traditional algorithms like Logistic Regression and Naïve Bayes performed poorly, with accuracies below 70%, highlighting their limitations for imbalanced medical datasets. While SMOTE and ADASYN produced slightly lower accuracy than random oversampling (86-91% versus 99%), they offered more sophisticated approaches to handling class imbalance by generating synthetic samples or focusing on difficult-to-classify cases.

- The study's results surpassed previous research efforts, which had achieved 77-88.7% accuracy, likely due to the comprehensive sampling strategies employed. However, the exceptionally high accuracy raises questions about potential overfitting to the specific dataset used, emphasizing the need for external validation across different patient populations and healthcare settings.

IV. Conclusion

The findings underscore the critical importance of addressing class imbalance in medical machine learning applications through appropriate sampling techniques. While SVM with oversampling achieved the highest performance metrics, Random Forest emerged as a more robust and potentially more practical choice for clinical implementation due to its consistent results across different sampling methods. The study demonstrates that careful preprocessing and algorithm selection can significantly improve heart disease prediction accuracy, but also highlights the need for further validation to ensure these models generalize well to diverse patient populations. Future research should focus on testing these approaches in multi-center studies, exploring hybrid techniques that combine different sampling methods with advanced algorithms, and developing explainability features to increase clinician confidence in the models' predictions. The results suggest promising directions for implementing machine learning in preventive cardiology, while also cautioning against over-optimism from single-study performance metrics.

References

[1]. J. Thomas and G. Y. Lip, "Novel risk markers and risk assessments for cardiovascular disease," Circulation Research, vol. 120, no. 1, pp. 133-149, 2017. DOI: 10.1161/CIRCRESAHA.116.309955.

[2]. M. Alaa et al., "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," PLOS ONE, vol. 14, no. 5, p. e0213653, May 2019. DOI: 10.1371/journal.pone.0213653.
[3]. S. F. Weng et al., "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PLOS ONE, Apr. 2017. DOI: 10.1371/journal.pone.0174944.
[4]. R. Nakanishi et al., "Machine learning in predicting coronary heart disease and cardiovascular disease events: Results from the Multi-Ethnic Study of Atherosclerosis (MESA)," Journal of the American College of Cardiology, vol. 71, no. 11, Mar. 2018.
[5]. S.M.Giri Rajkumar, S.Hari Prasath, G.Ragavantiran, I.Shanjith Kumar, S.Syed Athaullah Srimanti Roychoudhury, "Design of Plant Disease Identification System Using SVM Classifier", Journal of Next Generation Technology(ISSN: 2583-021X), 3(1), pp. 16-22. July 2023
[6]. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, 2019. DOI: 10.1109/ACCESS.2019.2923707.
[7]. Gavhane et al., "Prediction of heart disease using machine learning," in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275-1278.
[8]. M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," in 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEECT), 2017.
[9]. M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using K-nearest neighbor and genetic algorithm," Procedia Technology, vol. 10, pp. 85-94, 2013.
[10]. N. Al-milli, "Backpropogation neural network for prediction of heart disease," Journal of Theoretical and Applied Information Technology, vol. 56, no. 1, pp. 131-135, 2013.
[11]. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, "Machine learning techniques for heart disease prediction," International Journal of Scientific & Technology Research, vol. 8, no. 11, Nov. 2019.
[12]. Dr. B. Santhosh Kumar, T.Daniya, Dr. J.Ajayan, "Breast Cancer Prediction Using Machine Learning Algorithms", International Journal of Advanced Science and Technology, ISSN: 2005-4238, Vol. 29, no. 3, pp. 7819 - 7828, Mar.2020
[13]. Singh, A., Kumar, T. C. A., Mithun, T., Majji, S., Rajesh, M., & Anusha, P. (2021, December). Image Processing Approaches for Oral Cancer Detection in Color Images. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 817-821). IEEE.
[14]. Jahnavi, Y., Kumar, P. N., Anusha, P., & Prasad, M. S. (2022, November). Prediction and Evaluation of Cancer Using Machine Learning Techniques. In International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology (pp. 399-405). Singapore: Springer Nature Singapore.