



Article

Realistic Synthetic Electronic Health Record Data Generation for Cardiovascular Risk Prediction

Kavitha Bai A. S * and J. Somasekar

Department of Computer Science and Engineering, Faculty of Engineering and Technology,
JAIN (Deemed-to-be University), Bangalore 562112, Karnataka, India

* Correspondence: kavithabai.as@jainuniversity.ac.in

Highlights

What are the main findings?

- High-fidelity, privacy-preserving synthetic EHR data generated using TabDDPM with LLM-guided refinement.
- Knowledge-guided refinement improves clinical plausibility, data diversity, and downstream cardiovascular risk prediction performance.

What are the implications of the main findings?

- Provides a secure framework for AI research by enabling EHR data sharing without compromising patient privacy.
- Supports the development of reliable and generalizable cardiovascular risk prediction models and can be extended to other healthcare applications.

Abstract

Background: Increasing access to Electronic Health Records (EHRs) has enabled the development of Machine Learning (ML) models to predict early cardiovascular disease (CVD) risk. Nevertheless, real EHR data is usually limited in availability and distribution because of confidentiality issues, legal limitations, and insufficient access. Synthetic data generation has become a promising approach to overcome such difficulties. **Objectives:** This study proposes a Large Language Model (LLM)-guided retrieval-aware framework for generating realistic synthetic EHR data on a large scale to train ML models to predict cardiovascular risks accurately. **Methods:** The framework uses a Tabular Denoising Diffusion Probabilistic Model to learn the underlying distribution of the original dataset and generate an initial synthetic dataset. To improve the clinical plausibility of the generated data, a knowledge-guided refinement module with LLaMA 2 13B combined with Retrieval-Augmented Generation (RAG) is introduced. The LLM analyzes statistical trends in real and synthetic data while retrieving relevant medical information to detect and correct clinically implausible correlations. **Results:** Experimental findings show that the optimized synthetic dataset preserves important statistical features of the original data and achieves high performance in predicting CVD. **Conclusions:** Thus, the framework offers a scalable, privacy-conservative method for generating realistic synthetic healthcare data suitable for medical research.



Academic Editor: Raffaele Serra

Received: 19 June 2026

Revised: 16 July 2026

Accepted: 21 July 2026

Published: 25 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: cardiovascular diseases; diffusion model; large language model; machine learning; retrieval-augmented generation; sudden cardiac death

1. Introduction

Cardiovascular disease (CVD) is one of the major causes of morbidity and mortality worldwide, posing a huge burden to the healthcare systems [1]. Among its most severe outcomes is Sudden Cardiac Death (SCD) as it is often difficult to predict and is fatal, particularly among older adults [2,3]. Early diagnosis of individuals at high risk of SCD is essential for providing preventive measures as well as enhancing patient outcomes [4,5]. The World Health Organization (WHO) has highlighted risk prediction as a key action to achieve its objectives of minimizing mortality rates linked to CVD by 25% by the year 2025 [6]. The reliable risk prediction requires the availability of comprehensive and high-quality patient data in the form of Electronic Health Records (EHRs) that includes medical history, demographic, and lab results required to construct an accurate predictive model [7,8]. Nevertheless, real-world EHR data are often unavailable due to privacy policies, ethical considerations, and institutional demands that restrict access to such data and complicate the development and validation of efficient models [9]. Generative Artificial Intelligence (GenAI) has emerged as an innovative solution in this regard that allows creation of privacy-aware and realistic synthetic datasets [10,11]. By reproducing the statistical characteristics of actual EHRs, GenAI allows for the training of predictive models, data augmentation, and reduces biases arising from underrepresented populations or incomplete datasets, without compromising patient privacy [12,13]. Integrating synthetic data with actual EHRs can enhance cardiovascular risk forecasting, allow earlier detection of SCD, and support the development of Artificial Intelligence (AI)-based healthcare tools [14,15].

Several research works have investigated the issues of generative modeling to overcome the limitations associated with limited access to real EHR data [16]. For example, EHR-Safe, a generative modeling system, has been constructed to generate very realistic and privacy-aware synthetic EHR data using a two-stage architecture. It consists of sequential encoder–decoder networks and Generative Adversarial Networks (GANs) [17]. Similarly, a GAN-based framework called Electronic Health Records–Mixed-type–Generative Adversarial Network (EHR-M-GAN) has been developed for jointly combining heterogeneous time-series EHR data for AI applications. This model captured the multifaceted and spatio-temporal correlation present in trajectories of patients, while mimicking the nature of clinical decision-making [18]. In the framework of cardiovascular risk prediction, a hybrid Convolutional Neural Network (CNN)-GAN model has been developed to address the problem of imbalanced and limited medical datasets. The CNN module identifies meaningful clinical features and the GAN module creates realistic synthetic samples to supplement the dataset and enhance model robustness [19]. In addition to CVD risk prediction, identifying treatment-related complications such as anticoagulant-associated bleeding risk is essential for personalized clinical management. Previous studies have investigated risk factors for intracranial hemorrhage during anticoagulant therapy, emphasizing the need for reliable clinical prediction models and representative healthcare datasets [20]. Inflammatory processes also play an important role in cardiovascular risk development and long-term prognosis. Previous research on noninfectious thoracic aortitis has demonstrated the impact of chronic inflammation on cardiovascular outcomes, highlighting the importance of incorporating clinically relevant factors in predictive modeling [21].

Although these developments have been made in the prediction of CVDs using EHR data, several challenges remain: (i) limited access to high-quality data due to privacy, ethical and legal constraints and high economic costs which restrict data sharing for AI development [22]; (ii) the complex nature of EHR data and problems such as class imbalance and under-representation pose challenges to effective modeling [23,24]; (iii) existing methods for synthetic data generation tend to overfit the biases in the original datasets and lack diversity [25]; (iv) GAN-based models face stability issues during training and convergence

issues owing to the simultaneous training of the generator and discriminator [26]; and (v) GANs often generate limited or repetitive samples failing to capture the full diversity of the real dataset [27]. Therefore, the proposed model develops a knowledge-guided synthetic data generation framework to generate realistic EHR data. This data can be used for training Machine Learning (ML) models for improving the accuracy of CVD risk prediction, particularly for SCD. This paper presents the following novelties beyond the state of the art:

- **Synthetic EHR Data Generation Framework:** The proposed work develops a new framework that combines Tabular Denoising Diffusion Probabilistic Model (TabDDPM) with a large language model (LLM)-guided retrieval-aware mechanism for generating realistic synthetic EHR data to predict CVD risk.
- **Knowledge-Guided Refinement with LLM and RAG:** A knowledge-guided refinement system is proposed based on LLaMA 2 13B and Retrieval-Augmented Generation (RAG) to identify and remove clinically implausible patterns in the generated synthetic data to make it more clinically realistic and reliable.
- **Privacy-Preserving Data Utility in Healthcare Analytics:** The proposed framework allows the generation of synthetic healthcare data to ensure the privacy of patients and the development of effective ML-based predictive models for medical analytics.
- **Validation of Cardiovascular Risk Prediction:** The generated synthetic dataset is validated by statistical similarity and downstream performance evaluation of the ML to show its suitability in training predictive models for CVD risk.

The paper is structured as follows. The recent advancements in generating synthetic datasets are studied in Section 2. The suggested framework is detailed in Section 3, the experimental findings are described in Section 4, and Section 5 concludes the study.

2. Related Work

In this section, we analyze existing research on generative models for synthetic EHR data generation in healthcare. Kita et al. [28] introduced a diffusion model-based approach for generating synthetic EHR data and evaluated its utility in clustering disease patterns using Dirichlet Process Mixture Models (DPMMs). The model improved training stability and preserved complex feature dependencies in medical records. However, the model struggled to capture rare-disease patterns, limiting its ability to fully represent underrepresented clinical scenarios. To solve these challenges, Lu et al. [29] developed a Multi-label Time-series GAN (MTGAN) with Gated Recurrent Unit (GRU) for generating EHR and jointly improved the generation of rare-diseases-data quality. While MTGAN enhanced statistical representation for underrepresented conditions, it could not generate continuous real-valued features, reducing its applicability for fine-grained clinical measurements. Tian et al. [30] introduced a TIMEDIFF diffusion probabilistic model for producing varied synthetic EHR data using a Bidirectional Recurrent Neural Network (Bi-RNN). The model was evaluated using Medical Information Mart for Intensive Care III and IV (MIMIC-III/IV) and the electronic Intensive Care Unit (eICU) Collaborative Research Database. Nevertheless, the model produced varied synthetic data. It lacked systematic evaluation across different healthcare datasets, restricting its generalizability. Alqulaity et al. [31] introduced a Conditional GAN (CGAN) model for producing high-quality synthetic tabular data for CVD using data from King Faisal Specialist Hospital in Saudi Arabia. While the model effectively captured numerical characteristics of the actual data, its validation was limited to this single dataset, restricting generalizability to other medical populations and feature distributions.

To evaluate the synthetic tabular medical data, Kurakova et al. [32] developed a comprehensive evaluation framework that assesses synthetic data in four critical dimensions using Generative algorithms. The framework used Tabular Variational Autoencoder

(TVAE), Conditional Tabular GAN (CTGAN), and REaLTabFormer. The framework ensures that the synthetic data is sufficiently similar to the real data while preserving confidentiality. Jia et al. [33] developed a novel Gaussian diffusion-based Medical Service Data (MSDGD) creation model to effectively capture the randomness of medical data using a U-Net model. This enabled the generation of more realistic and representative synthetic datasets, while providing privacy through enhanced generalizability of the model. However, the model faced challenges in balancing computational efficiency with high-fidelity data generation, indicating a trade-off between resource usage and data quality. Thus, by analyzing existing models, we identified some issues in real-world EHR data that are often unavailable because of privacy and disproportional and limited characterization of rare-disease distributions, which impede the creation and validation of robust ML models. Current generative methods, including GANs, Variational Autoencoder (VAEs), and diffusion models, are either unstable to train, lack diversity, or do not guarantee clinical plausibility, especially for high-risk patients. Therefore, there is a need to create a privacy-preserving, realistic, and clinically accurate synthetic EHR generation framework that allows for effective CVD risk modeling while addressing the constraints of existing approaches.

3. Proposed Methodology

The proposed methodology develops a framework for generating realistic synthetic EHR data to train ML models for accurate CVD risk prediction. The framework initially preprocesses the original EHR dataset and trains a TabDDPM diffusion model for learning the underlying data distribution and generating an initial synthetic dataset. A pre-trained LLaMA 2 13B model, augmented with RAG, is then used as a knowledge-guided reasoning module to compare real and synthetic data statistics and identify clinically implausible patterns.

Based on the feedback generated by the LLM, constraint-based refinements are applied to the diffusion sampling process to improve realism. The refined synthetic data is finally used to train downstream ML models which are trained on a real dataset to assess the effectiveness and reliability of the framework. The overall architecture is depicted in Figure 1.

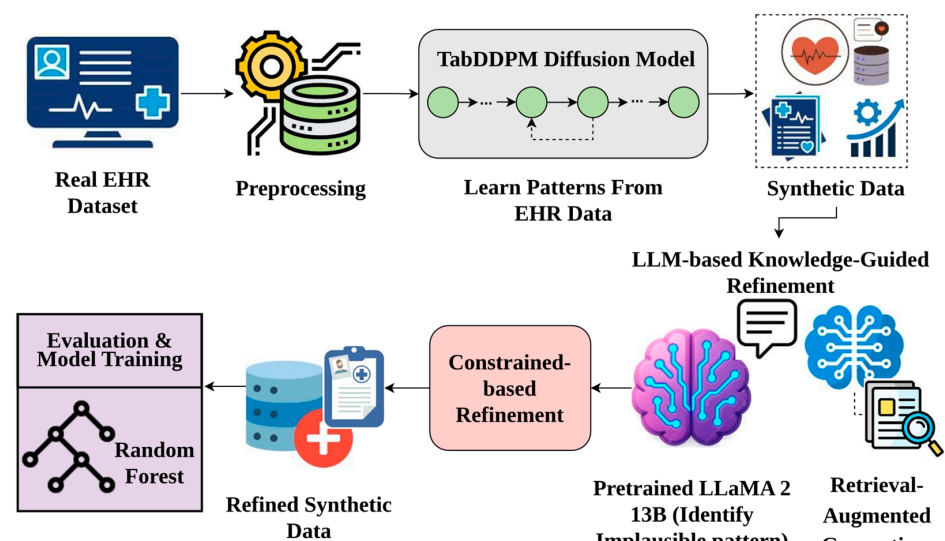


Figure 1. Overall proposed framework for knowledge-guided synthetic EHR data generation using TabDDPM diffusion model and LLaMA 2 13B-based Retrieval-Augmented Generation. The framework consists of EHR data preprocessing, TabDDPM-based synthetic data generation, LLaMA 2 13B with RAG-based knowledge-guided refinement, and downstream cardiovascular risk prediction evaluation.

3.1. Dataset Description

The proposed work utilizes the University of California, Irvine (UCI) Heart Disease Dataset [34], which is a multivariate dataset comprising multiple clinical attributes used for statistical and ML analysis. This dataset includes 14 features, including age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate attained, angina brought on by exercise, ST depression brought on by exercise in comparison to rest, the slope of the peak exercise ST segment, number of major vessels, and thalassemia. This dataset is commonly utilized for predicting whether that specific individual has heart disease or not, based on specified attributes of a patient. This helps to derive some meaningful insights that contribute to a better understanding of cardiovascular risk factors.

3.2. Data Preprocessing

The dataset is loaded, and then the target variable is converted into a binary classification label. Categorical attributes are encoded using LabelEncoder [35] to transform the categorical attributes into numerical representations. Finally, missing values in the dataset are handled using median imputation [36] to ensure consistency and stability during the model training.

3.3. Synthetic Data Generation

To generate realistic synthetic EHR data, the proposed framework uses a TabDDPM [37]. Diffusion models are trained to match the original distribution of the data by progressively adding noise to the original data using a forward diffusion process and learning to reconstruct using a reverse denoising process, as shown in Figure 2. This approach provides more stable training and improved diversity than GAN-based methods. In the proposed work, the EHR dataset contains both categorical and numerical clinical attributes. A tabular sample x is denoted by Equation (1):

$$x = [x_{num}, x_{C_1}, x_{C_2}, \dots, x_{C_c}] \quad (1)$$

where x_{num} represents numerical features, and x_{C_i} are categorical attributes. Numerical attributes are modeled using Gaussian diffusion, and categorical attributes are modeled using multinomial diffusion. All categorical variables are transformed into one-hot encoding vectors, and the numerical variables are normalized by the use of a Gaussian quantile transformation to maintain a uniform data distribution. During the forward diffusion process, noise is accumulated in the data across multiple steps until the distribution of the data is close to random noise. Each of the categorical features is diffused independently to maintain the feature-level independence while preserving the overall data structure. The reverse diffusion process is modeled using a multi-layer neural network that reconstructs the original data from noisy inputs by predicting the noise for numerical attributes and the categorical distributions for categorical features. The TabDDPM model is trained with a combined loss function $\mathcal{L}_{TabDDPM}$, as illustrated in Equation (2):

$$\mathcal{L}_{TabDDPM} = \mathcal{L}_{simple} + \frac{1}{C} \sum_{i=1}^C \mathcal{L}_i \quad (2)$$

where $\mathcal{L}_{simple}$ is the Mean Squared Error (MSE) loss for numerical features, $\mathcal{L}_i$ is the Kullback–Leibler (KL) divergence associated with the i -th categorical diffusion process, and C is the total number of categorical diffusion processes. After training, the model produces initial synthetic EHR data that retains the numerical properties of the actual data. This generated dataset is then optimized with the suggested LLM-guided retrieval-aware

knowledge-refinement module to enhance clinical realism and usefulness to downstream predictive models of CVDs.

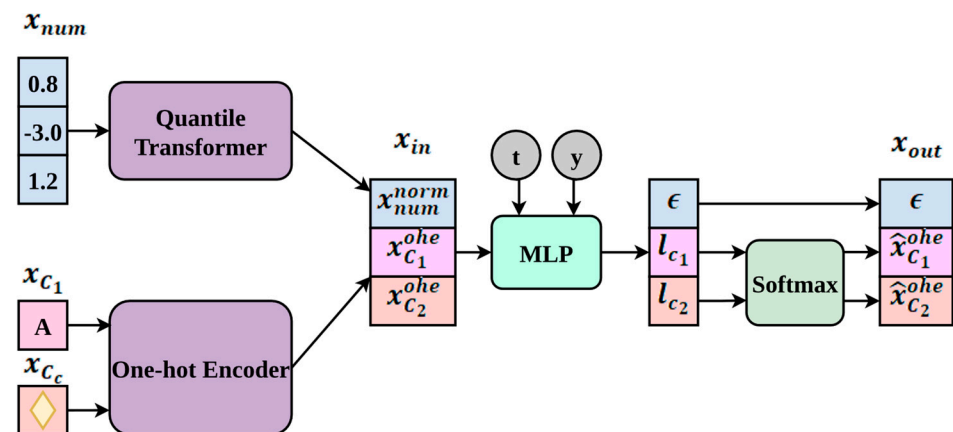


Figure 2. Architecture of tabular denoising diffusion probabilistic model for learning patterns from real Electronic Health Record data. Numerical features are modeled using Gaussian diffusion, while categorical features are processed using multinomial diffusion with a combined MSE and KL divergence optimization objective.

3.4. Knowledge-Guided Refinement

Although the TabDDPM diffusion model learns the distribution of the original EHR dataset, it generates synthetic records with clinically inconsistent attribute combinations or relations that are not fully captured by statistical similarity measures. To address this limitation, the proposed framework uses a knowledge-guided refinement module based on a pretrained LLaMA 2 13B large language model integrated with RAG [38,39]. The LLM performs clinical reasoning by analyzing relationships among medical attributes, while the RAG module provides external cardiovascular knowledge for evidence-based refinement. The overall refinement process is illustrated in Figure 3.

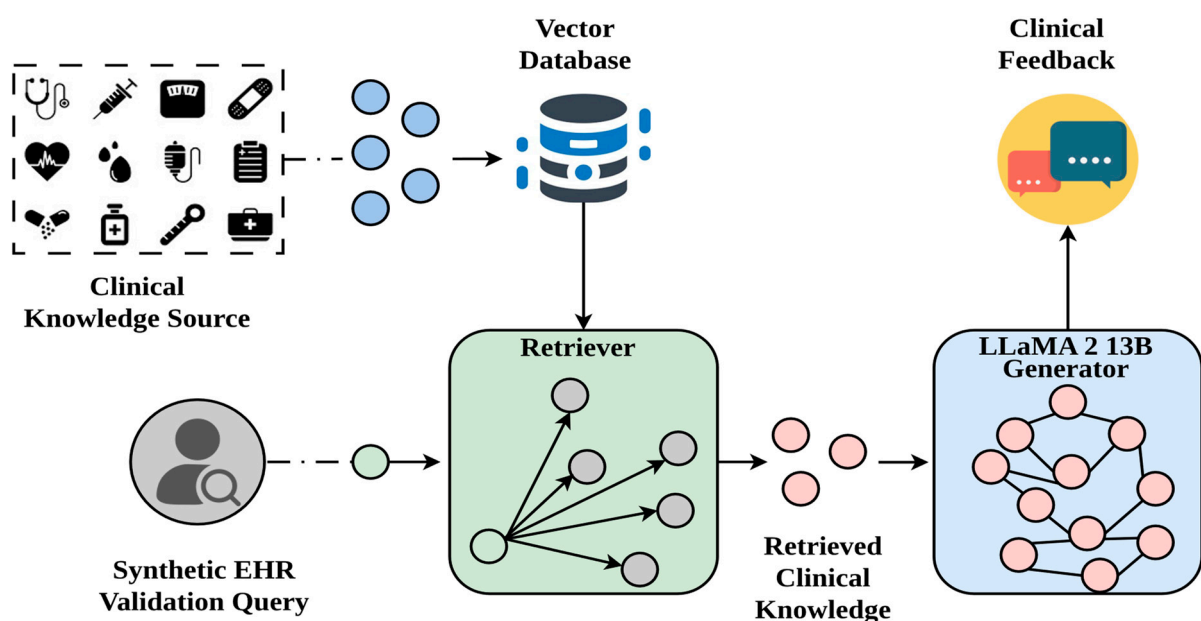


Figure 3. Architecture of the LLaMA 2 13B retrieval-augmented generation framework for knowledge-guided clinical validation. The module integrates clinical knowledge retrieval using FAISS with embedding-based similarity search and generates refinement constraints for improving the clinical consistency of TabDDPM-generated synthetic EHR records.

3.4.1. Knowledge Retrieval Using Retrieval-Augmented Generation

The refinement process begins after generating the initial synthetic EHR dataset using TabDDPM. Statistical summaries are obtained from both real and synthetic datasets such as descriptive statistics, feature distributions, correlation coefficients and ranges of acceptable clinical value. These summaries are combined with external medical knowledge retrieved through the RAG framework from cardiovascular guidelines, peer-reviewed literature and medical knowledge repositories. The information retrieved helps the LLaMA 2 model to assess the generated synthetic records with clinically validated relationships, diagnostic criteria and physiological ranges.

3.4.2. Prompt Construction and Clinical Reasoning

The retrieved knowledge and statistical summaries are compiled into a structured prompt fed to the LLaMA 2 model. The prompt consists of three parts: (i) statistical features of the original EHR data set, (ii) statistical features of the TabDDPM-generated synthetic data set, and (iii) retrieved cardiovascular knowledge describing clinical relationships between risk factors. The LLaMA 2 model compares these inputs to identify abnormal distributions, unrealistic correlations, contradictory attribute combinations, and medically inconsistent records. It then generates refinement recommendations and constraint rules based on this comparison. The prompt used in the proposed framework is organized as follows:

Input:

- Feature distributions, descriptive statistics, correlation coefficients and clinical ranges for the real EHR data set.
- Statistics of the created synthetic EHR dataset.
- Gained cardiovascular clinical knowledge from the RAG module.

Instructions: Using the retrieved cardiovascular clinical knowledge, compare the statistical properties of the real and the synthetic data sets. Detect anomalous combinations of medical attributes, unusual correlations and impossible relationships between features. Develop a set of refinement constraints that enhance clinical plausibility of the synthetic data while maintaining its similarity to the clinical data in the EHR.

Expected Output:

- Found discrepancies in the artificial dataset.
- Medical explanations retrieved from the literature that provided a clinical explanation.
- Recommended refinement constraints.

3.4.3. Clinical Constraint Generation

The refinement constraints generated by the LLaMA 2 model capture the medically consistent relationships between cardiovascular variables. These constraints are divided into four types: range constraints, logical consistency constraints, correlation constraints and distributional constraints. Range constraints ensure numerical variables remain within physiological scopes while logical consistency constraints ensure valid relations among clinical attributes. Correlation constraints preserve meaningful dependencies among cardiovascular risk factors and distributional constraints reduce the deviations between the real and synthetic data distributions.

3.4.4. Constraint-Based Synthetic Data Refinement

The generated constraints are then applied to the initial synthetic dataset to find and correct records with clinically inconsistent patterns. Each synthetic record is compared against the constraint rules and only the affected attributes are corrected, while the rest of the record's characteristics are maintained. After refinement, the updated synthetic dataset

is refitted with statistical summaries to check for consistency with the original data distribution. This iterative process of refinement continues until the pre-specified refinement criteria are satisfied. The detailed knowledge-guided refinement procedure is presented in Algorithm 1. The final refined synthetic dataset combines the distribution-learning power of TabDDPM with clinically based corrections from the LLaMA 2-RAG module, leading to an improved quality for downstream cardiovascular risk prediction models.

Algorithm 1. Knowledge-Guided Refinement Procedure

Input: Original EHR dataset D_{real} , initial synthetic dataset D_{sync} generated by TabDDPM

Output: synthetic dataset $D_{refined}$

Step 1: Create the synthetic data set of EHRs from the trained TabDDPM diffusion model.

Step 2: Extract statistical summaries from both the original and synthetic datasets such as feature distributions, descriptive statistics, correlation coefficients and clinical value ranges.

Step 3: Obtain relevant cardiovascular medical information from external clinical sources using the RAG framework.

Step 4: Create a structured prompt using real dataset statistics, synthetic dataset statistics and retrieved cardiovascular clinical information.

Step 5: Pass the prompt to the pretrained LLaMA 2 13B model.

Step 6: Conduct comparative clinical reasoning to detect clinically improbable values, abnormal relationships of features and inconsistent combinations of medical attributes.

Step 7: Compute refinement constraints from the clinical knowledge retrieved and the statistical comparison.

Step 8: Each synthetic record is compared to the clinical constraints generated.

Step 9: Refine records that violate one or more constraints but maintain the learned statistical properties.

Step 10: Re-run the synthetic data with the refined dataset to ensure statistical consistency and clinical plausibility.

Step 11: Output the final synthetic, refined EHR data for downstream CVD predictions.

3.5. Synthetic Data Evaluation and Model Training

After the knowledge-guided refinement process, the superiority and usability of the produced synthetic dataset are evaluated to ensure that it replicates the numerical properties of the actual EHR data. The original UCI Heart Disease dataset is randomly partitioned into 70% training, 10% validation and 20% testing subsets. The training set is used for TabDDPM training, synthetic EHR data generation and downstream ML model training. The validation set is utilized for hyperparameter optimization and model selection, whereas the testing set is reserved exclusively for the final performance evaluation. To prevent data leakage, all preprocessing operations are performed exclusively on the training data. Missing value imputation, categorical encoding, normalization and feature scaling parameters are estimated using only the training set and subsequently applied to the validation and testing sets. Synthetic EHR data generation using the TabDDPM model is conducted solely on the training partition, ensuring that no information from the validation or testing data influences the synthetic data generation process or the downstream model training. Hyperparameter optimization is performed using the validation set, while the testing set remains completely unseen until the final evaluation.

The analysis is aimed at measuring both the statistical similarity and predictive performance of the generated dataset. First, the statistical characteristics of the refined synthetic

data are compared with those of the real dataset by analyzing feature distributions, variable correlations, and clinically acceptable value ranges to determine whether the generated data preserves the significant statistical patterns of the original EHR data. The usefulness of the synthetic dataset in downstream ML tasks is subsequently evaluated by training a Random Forest (RF) classifier using the refined synthetic dataset for CVD risk prediction. The trained model is then evaluated using the unseen real testing dataset to determine whether the synthetic data effectively preserves clinically meaningful patterns required for reliable prediction. This evaluation assesses both the statistical consistency of the generated synthetic dataset and its ability to support robust and clinically meaningful CVD risk prediction.

4. Results and Discussion

This section analyzes the efficiency of the proposed framework for generating a synthetic dataset. This section discusses the proposed framework by analyzing the model performance using different metrics. The parameters used in the experimental analysis are summarized in Table 1.

Table 1. The parameters involved in the proposed knowledge-guided data generation framework.

Module	Parameter	Value
System Configuration	Programming Language	Python 3.11
	Deep Learning Framework	PyTorch 2.3
	CUDA Toolkit	CUDA 12.1
	Operating System	Windows 10 Pro 64-bit
	CPU	Intel® Xeon® E5-2680 v4 @ 2.40 GHz
	GPU	NVIDIA Quadro M2000 (4 GB VRAM)
	System RAM	16 GB
	Storage	480 GB SSD (WD Green) + 932 GB HDD
General Settings	Random Seed (NumPy)	42
	Random Seed (PyTorch)	42
	Batch Size	256
TabDDPM Diffusion Model	Training Epochs	200
	Diffusion Timesteps	1000
	Reverse Diffusion Steps	1000
	Optimizer	Adam
	Learning Rate	0.001
	Activation Function	ReLU
	Loss Function	MSE + KL Divergence
Synthetic Data Generation	Initial Latent Noise	Gaussian
	Synthetic Sample Size	Same as the original dataset
Knowledge-Guided Refinement	Large Language Model	LLaMA-2-13B-Chat
	Embedding Model	all-MiniLM-L6-v2
	Vector Database	FAISS

Table 1. *Cont.*

Module	Parameter	Value
Knowledge-Guided Refinement	Retrieval Top-k	5
	Prompt Temperature	0.2
	Maximum Output Tokens	1024

The proposed framework implementation was done using Python 3.11 and the deep learning framework PyTorch 2.3. We trained the TabDDPM diffusion model for 200 epochs with the Adam optimizer, a learning rate of 0.001 and a batch size of 256. The diffusion process had 1000 forward and reverse timesteps and we optimized the numerical and categorical features with a mixture of MSE and KL divergence loss. For knowledge-guided refinement, we integrated the pretrained LLaMA-2-13B-Chat model with a RAG framework. Clinical knowledge retrieval was performed using Facebook AI Similarity Search (FAISS), where the all-MiniLM-L6-v2 embedding model encoded the knowledge corpus and the top five most relevant documents were retrieved based on cosine similarity. The language model generated the refinement constraints with a prompt temperature of 0.2 and maximum output length of 1024 tokens. All experiments were conducted with a fixed random seed of 42 for reproducibility purposes.

4.1. Performance Evaluation of the Proposed Knowledge-Guided Synthetic Data Generation Framework

4.1.1. Predictive Performance and Statistical Similarity Analysis

Table 2 shows the comparison of predictive performance and statistical resemblance between the real EHR dataset and the generated realistic dataset. Classification metrics including Accuracy, Precision, Recall, F1-Score and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) indicate that the models trained on synthetic data perform similarly to the model trained on original data and exhibit high predictive power. The high classification performance is attributed to the TabDDPM-based synthetic data generation coupled with the LLM-guided knowledge refinement, which preserves both the statistical properties and clinically relevant relationships of the original EHR data. The distributional similarity of the datasets is assessed using statistical measures like KL Divergence, Jensen–Shannon (JS) Divergence, Wasserstein Distance and Correlation Difference. The minimal divergence and proximity scores confirm that the synthetic data retains important statistical features associated with the original dataset without compromising privacy.

Table 2. Performance and statistical similarity comparison between original and synthetic EHR data.

Metric	Original Data Value	Synthetic Realistic Data Value
Accuracy	99.02%	95.14%
Precision	99.31%	96.38%
Recall	98.76%	94.05%
F1-Score	99.03%	95.20%
ROC-AUC	99.54%	97.34%
KL Divergence	0.021	0.055
JS Divergence	0.014	0.037
Wasserstein Distance	0.412	0.686
Correlation Difference	0.018	0.043

This shows the effectiveness of the proposed knowledge-guided and diffusion-based refinement architecture in generating realistic, statistically consistent, and clinically meaningful synthetic EHR data.

4.1.2. Clinical Feature Importance Analysis for Cardiovascular Risk Prediction

Table 3 shows the ranking of clinical parameters based on their contribution to cardiovascular risk prediction by the trained RF classifier. The importance scores indicate the influence of each feature on the prediction result: the higher the score, the greater the influence for risk classification. The analysis indicates that the most important parameters related to increased cardiovascular risk are the number of major vessels (CA), ST depression (Oldpeak), and abnormalities related to thalassemia. Other clinical factors, such as chest pain type, maximal Heart Rate (HR), exercise-induced angina, age and cholesterol levels, also influence risk assessment. The forecasted result is the binary status of CVD (present or absent) that was used to assess the model. The results improve the interpretability of the proposed framework by connecting decisions made by the model with clinically relevant risk factors.

Table 3. Clinical feature importance and association with cardiovascular risk prediction using the proposed framework.

Rank	Clinical Feature	Importance Score	Association with Cardiovascular Risk
1	Number of Major Vessels (CA)	0.1847	A greater number of affected vessels is strongly associated with increased cardiovascular risk.
2	ST Depression (Oldpeak)	0.1619	Higher ST depression reflects greater myocardial ischemia and elevated risk.
3	Thalassemia (Thal)	0.1386	Abnormal thalassemia findings are associated with increased cardiovascular risk.
4	Chest Pain Type (CP)	0.1174	Asymptomatic and atypical chest pain patterns contribute substantially to risk prediction.
5	Maximum Heart Rate (Thalach)	0.0958	Lower maximum heart rate achieved during exercise is associated with higher risk.
6	Exercise-Induced Angina (Exang)	0.0736	Presence of exercise-induced angina indicates elevated cardiovascular risk.
7	Age	0.0584	Increasing age contributes to a higher probability of CVD.
8	Cholesterol (Chol)	0.0449	Elevated serum cholesterol moderately increases CVD.

4.1.3. Effectiveness of Knowledge-Guided Refinement

Figure 4 shows that the synthetic EHR data quality gradually increases with the increasing number of knowledge-based refinement iterations using a LLaMA 2 13B model with RAG. The x-axis shows the refinement iteration, and the y-axis shows the data quality score as a percentage. Initially, the diffusion-generated dataset is of quality 65% at iteration 0. With each refinement iteration led by the LLM and retrieved clinical knowledge, the score improves gradually up to a 94% score by the fifth iteration, approaching the 95% target threshold. This shows the effectiveness of using iterative LLM-guided feedback to increase the realism, consistency, and clinical plausibility of a synthetic EHR dataset.

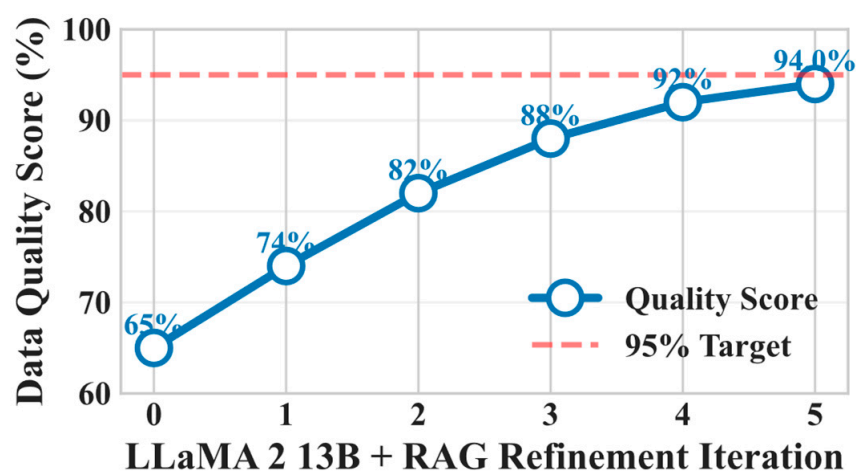


Figure 4. Improvement in synthetic EHR data quality across iterative knowledge-guided refinement using LLaMA 2 13B with RAG on TabDDPM-generated samples. Data quality score (%) is calculated based on clinical consistency and statistical similarity measures across five refinement iterations.

Figure 5 provides a comparison between the predictive performance of the ML models trained on generated synthetic EHR data without refinement and with LLaMA 2 13B and RAG knowledge-guided refinement framework. The x-axis shows the refinement method and the y-axis shows the model accuracy (%). Without LLM refinement, the mean accuracy is 75.2%, with wider variability indicating inconsistent data quality. Conversely, using LLaMA 2 13B with RAG yields a higher average accuracy of 94.7% with a narrower distribution, demonstrating that knowledge-guided refinement improves the realism, consistency, and downstream predictive performance of the generated synthetic EHR data.

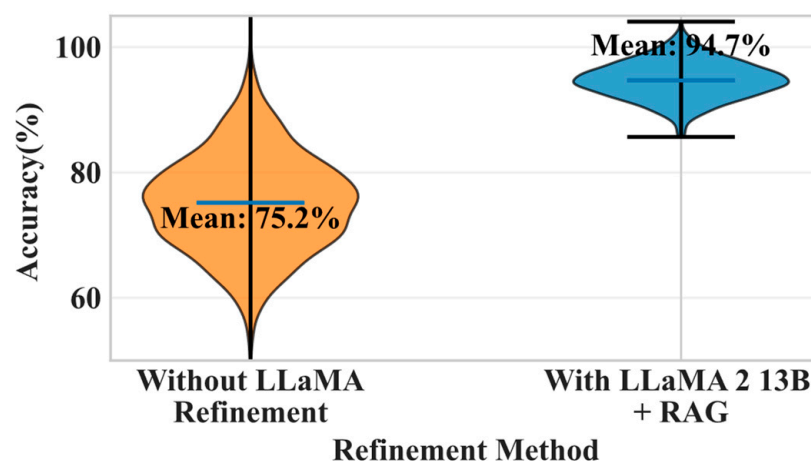


Figure 5. Comparison of predictive accuracy using synthetic EHR data with and without LLaMA 2 13B and RAG knowledge-guided refinement. The accuracy (%) is obtained using the Random Forest classifier evaluated on the unseen real EHR test dataset to measure the effect of refinement on downstream cardiovascular risk prediction.

4.1.4. Distributional Similarity Between Real and Synthetic Data

Figure 6 compares real and synthetic EHR datasets using three statistical distance measures across multiple clinical features. In Figure 6a, the values of the KL divergence are below the predefined threshold in most instances, indicating that the probability distribution of the synthetic data closely matches the real data. In Figure 6b, the values of the JS divergence are relatively low and remain below the threshold of 0.05 for most variables, demonstrating consistent similarity between the two datasets.

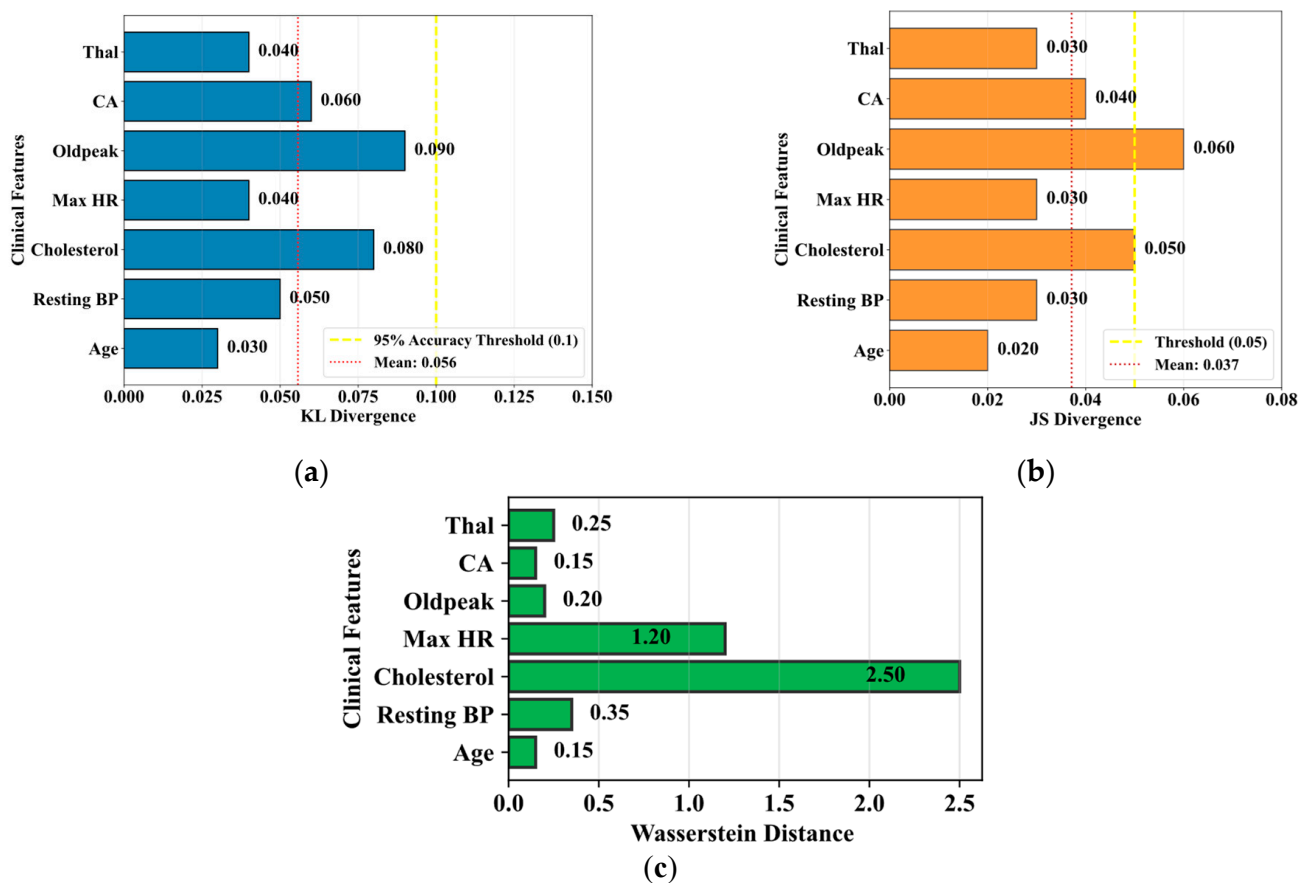


Figure 6. Distribution similarity evaluation between real and synthetic EHR data using statistical divergence metrics: (a) KL divergence across clinical features; (b) JS divergence across clinical features; and (c) Wasserstein distance across clinical features. The metrics are computed across clinical features of the EHR dataset, where lower values indicate closer distributional similarity between real and synthetic data.

In Figure 6c, the Wasserstein distance quantifies the discrepancies in the distributions based on the magnitude and dispersion of data. Where most features have small values, cholesterol and maximum heart rate have slightly larger values. Altogether, the results suggest that the generated synthetic data are capable of preserving statistical aspects and properties of the original EHR data.

Figure 7 shows the relationship between theoretical quantiles and the ordered values of both the real and synthetic datasets to measure how well the generated dataset maintains statistical properties of the original data. Theoretical quantiles represent expected values of some reference statistical distributions, and ordered values represent actually observed values arranged in a rising sequence. In Figure 7a, the actual dataset represents a set of data points that closely align with the fitted reference line. This shows that actual values are distributed in accordance with the anticipated statistical data with negligible variation at the extreme values. In Figure 7b, the generated synthetic dataset with the proposed diffusion-based framework demonstrates a very similar behavior to the fitted line, which means that the synthetic data effectively represent the underlying distribution of the real EHR data. The similarity between the two datasets implies that there is a high degree of distributional consistency, indicating that the proposed framework can generate realistic synthetic data that can be used for ML analysis.

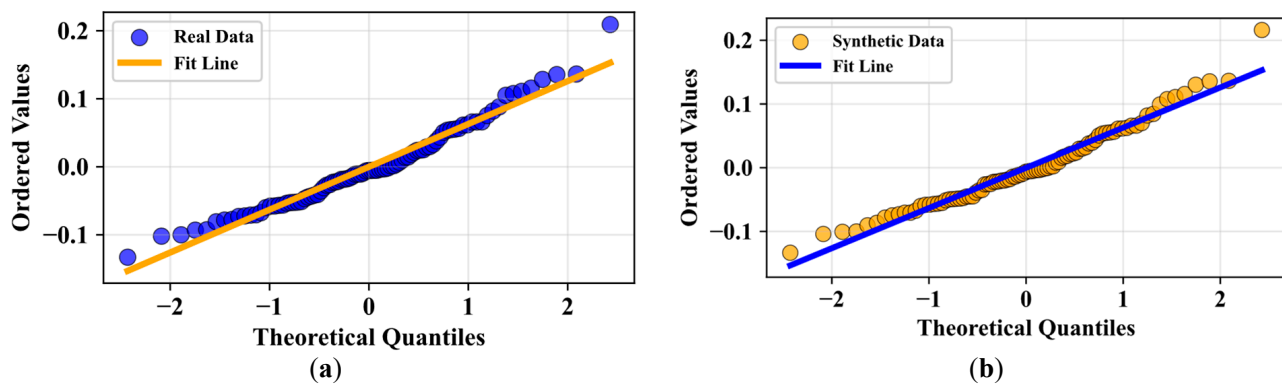


Figure 7. Distributional alignment between (a) real EHR dataset and (b) synthetic EHR dataset generated by the proposed TabDDPM + LLaMA 2 + RAG framework for validating statistical fidelity of the generated dataset. Quantile–quantile analysis is performed to evaluate the preservation of feature distributions between the original and generated datasets.

4.1.5. Privacy Preservation Evaluation

To evaluate the privacy-preserving capability of the proposed framework, dedicated privacy metrics were computed in addition to the statistical similarity analysis. Table 4 summarizes the results for Membership Inference Attack (MIA) Success Rate, MIA AUC, Attribute Disclosure Risk, Re-identification Risk, Nearest Neighbor Distance Ratio (NNDR), Distance to Closest Record (DCR), Privacy Leakage Score and Privacy Preservation Score.

Table 4. Privacy evaluation of the proposed synthetic EHR data generation framework.

Privacy Metric	Value
Membership Inference Attack Success Rate (%)	51.84
Membership Inference Attack AUC	0.517
Attribute Disclosure Risk (%)	3.42
Re-identification Risk (%)	2.08
Nearest Neighbor Distance Ratio (NNDR)	1.91
Distance to Closest Record (DCR)	0.842
Privacy Leakage Score	0.028
Privacy Preservation Score (%)	97.2

The results indicate a low disclosure risk and limited susceptibility to privacy attacks while maintaining high data utility. The results demonstrate that the generated synthetic EHR data are well-balanced in terms of privacy preservation and downstream analytical usefulness.

4.1.6. Correlation and Feature Dependency Preservation

Figure 8 compares the distribution of the pairwise correlation coefficients between the features of the real EHR dataset (blue) and the generated synthetic dataset (orange). The x-axis is the correlation coefficient, which shows the strength and direction associated with variables, and the y-axis is the frequency, which is the count of feature pairs that are within a given range of correlation. The high overlap between the two distributions indicates that the generated data is applicable in ensuring that the statistical relationship of the original data is maintained. This shows that the diffusion-based generation and knowledge-informed refinement strategy preserves significant feature dependencies, demonstrating superior

distributional fidelity and reflecting whether the generated dataset trains downstream ML models.

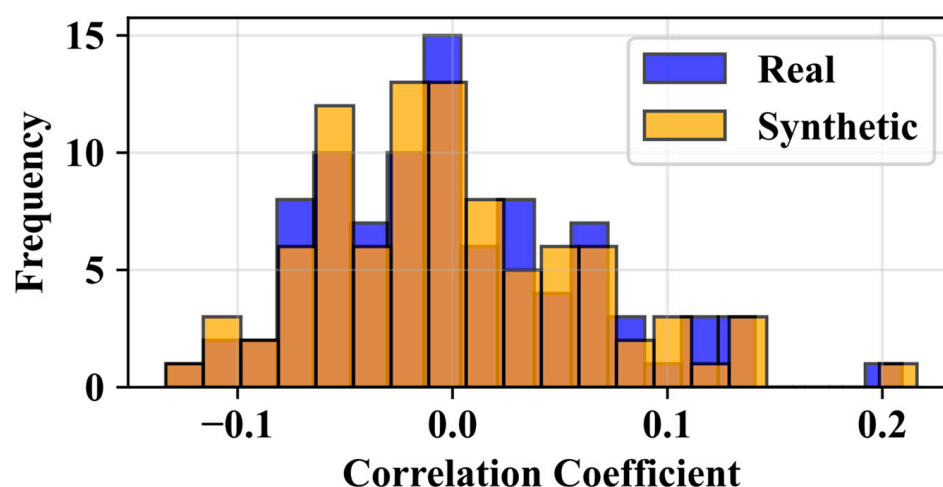


Figure 8. Comparison of pairwise feature correlation coefficient distributions between real and synthetic EHR data. Correlation coefficients among clinical variables are calculated to evaluate the preservation of feature dependencies and relationships after synthetic data generation. The dark brown regions represent the overlapping portions of the real and synthetic histograms, indicating that both datasets exhibit similar correlation coefficient distributions.

Figure 9 compares the correlation coefficients of selected feature pairs between the generated real EHR dataset (blue) and the generated synthetic dataset (orange). The feature pairs are provided on the x-axis, and the correlation coefficient on the y-axis shows how strong and in which direction the association between the two variables is. When the value is positive, it implies a direct relationship, and when the value is negative, this implies an inverse relationship. The close resemblance between the actual and synthetic bars across most feature pairs implies that the generated data efficiently maintains the relationships between clinical variables. This demonstrates the ability of the diffusion-based generation with knowledge-directed refinement architecture to preserve realistic data dependencies.

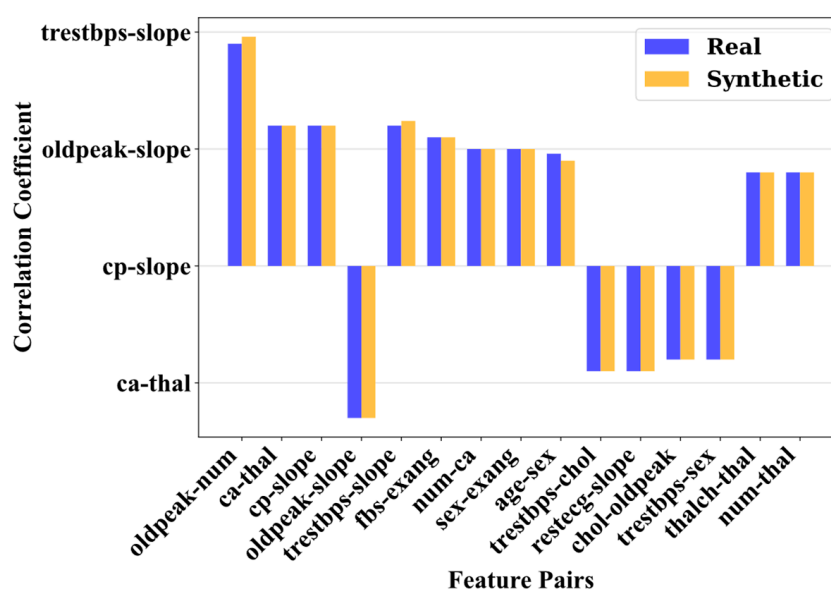


Figure 9. Comparison of selected pairwise feature correlation coefficients between real and synthetic EHR data. Correlation values of representative clinical feature pairs are compared to assess the preservation of linear relationships between variables.

Figure 10 compares the correlation between significant clinical variables in both the real and synthetic EHR datasets to assess how well the generated dataset contains underlying patterns. In Figure 10a–d, both datasets show a similar distribution, and overlapping data points show that the generated data maintains the overall structure of the actual data. Correlation coefficients of each pair of variables are also similar in real and synthetic datasets, with only slight differences. This similarity implies that the generated synthetic data effectively captures statistical associations among clinical variables. This shows that the suggested framework effectively maintains significant statistical relationships while preserving the privacy necessary for effective downstream ML analysis.

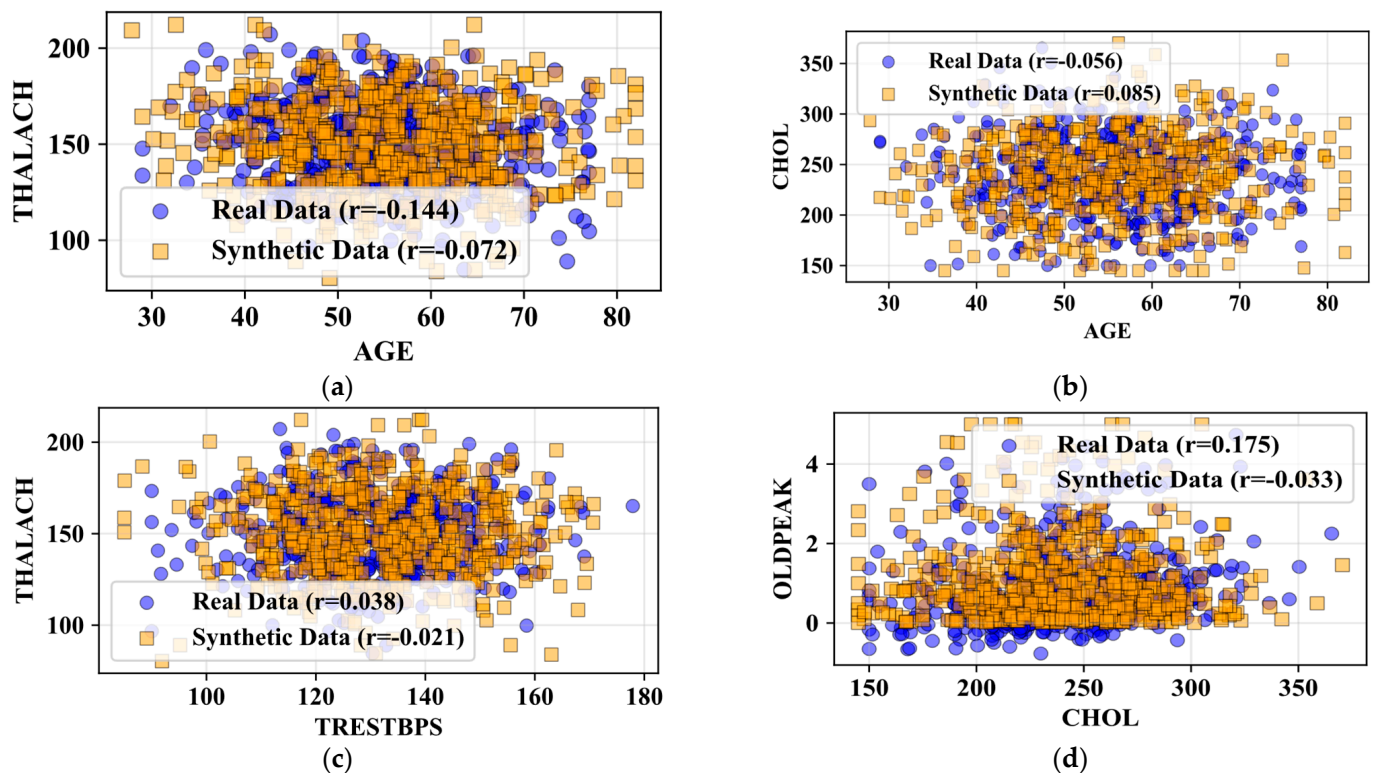


Figure 10. Relationship comparison between real and synthetic EHR data variables: (a) AGE vs. THALACH, (b) AGE vs. CHOL, (c) TRESTBPS vs. THALACH, and (d) CHOL vs. OLDPEAK. Scatter distributions and correlation coefficients are analyzed to evaluate the preservation of clinically relevant variable associations.

4.1.7. Diffusion Model Training Stability

Figure 11 indicates the development of training and validation diffusion loss in the form of MSE across 50 epochs throughout the TabDDPM diffusion model training to produce synthetic EHR data. The values of both curves indicate that as the number of epochs increases, the loss value decreases steadily, which indicates that the model is effectively learning the underlying distribution of the actual EHR data.

The short distance separation between the training and validation loss curves indicates high generalization and less overfitting. Convergence of models across epochs indicates the stability of loss, and it confirms the reliability of the trained diffusion model for generating realistic and high-quality synthetic healthcare data.

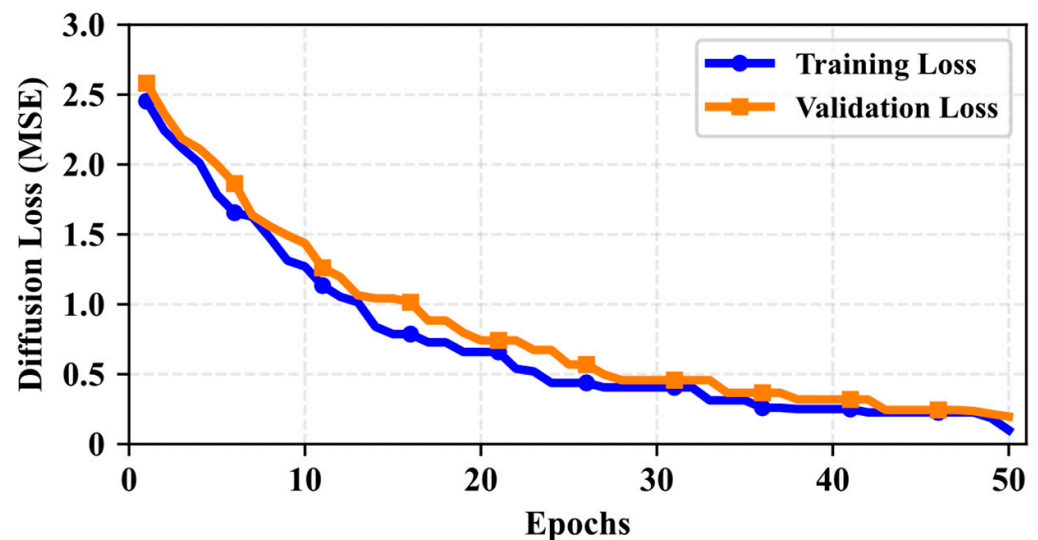


Figure 11. Convergence of training and validation diffusion loss during TabDDPM model training for synthetic EHR data generation. The diffusion model is trained for 200 epochs using a combined Mean Squared Error (MSE) and Kullback–Leibler (KL) divergence loss to optimize numerical and categorical feature reconstruction.

4.1.8. Five-Fold Cross-Validation Performance

To evaluate the robustness and generalization ability of the proposed framework, we performed five-fold cross-validation on the training data. Table 5 reports the classification performance for accuracy, precision, recall, F1-score and ROC-AUC across the five folds. The results show consistent performance and low standard deviations for all the evaluation metrics, indicating stable learning behavior and low sensitivity to data partitioning. These results indicate that the proposed framework generalizes well across different data subsets and does not depend on a single train-test split to achieve high predictive performance.

Table 5. Five-fold cross-validation performance of the proposed framework for cardiovascular disease risk prediction.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Fold 1	94.82	96.05	93.68	94.85	97.08
Fold 2	95.37	96.71	94.42	95.55	97.52
Fold 3	95.21	96.42	94.09	95.24	97.33
Fold 4	95.46	96.58	94.38	95.47	97.61
Fold 5	94.84	96.16	93.7	94.9	97.16
Mean $\pm$ SD	95.14 $\pm$ 0.30	96.38 $\pm$ 0.28	94.05 $\pm$ 0.36	95.20 $\pm$ 0.33	97.34 $\pm$ 0.23

4.1.9. Generalization Performance Across Multiple CVD Datasets

To prove the generalization capability of the proposed framework, experiments were conducted on multiple cardiovascular datasets including the Sudden Cardiac Death Database (SDDDB) [40], MUSIC [41], and the CVD dataset [42]. The SDDDB from PhysioNet is a specialized dataset developed to support research on sudden cardiac death and ventricular arrhythmias. It contains 23 complete Holter ECG recordings of patients who had sustained ventricular tachyarrhythmias or cardiac arrest. The recordings were mostly made in hospitals in Boston during the 1980s. The dataset consists of sinus rhythm, atrial fibrillation, and paced rhythms, offering cardiac patterns. The MUSIC dataset is a multicenter longitudinal, prospective database that predicts cardiac mortality and SCD among patients

with Chronic Heart Failure (CHF). It consists of 992 patients who were recruited in eight Spanish hospitals in the years 2003 to 2004 and had a mean follow-up of 44 months. The dataset includes clinical, demographic, laboratory, echocardiographic, and ECG data (with Holter records), as well as 24 h echocardiographic and high-resolution ECG data.

The CVD dataset has 11 characteristics with a target and 70,000 patient records. Factual information, medical examination results, and patient-provided information are the three categories of input features. Age, height, weight, gender, systolic and diastolic blood pressure, cholesterol, glucose, smoking, alcohol consumption, physical activity, and the presence or absence of CVD were all recorded at the time of the medical examination. Table 6 measures the effectiveness and generalization capability of the presented framework in generating synthetic EHR data across multiple CVD datasets. For each dataset, models trained using synthetic data attain performances that are similar to models trained on original data in Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. The suggested framework shows the best performance, confirming its capability to generate realistic, statistically consistent, and clinically meaningful synthetic datasets applicable for a wide range of medical data sources.

Table 6. Generalization performance of the proposed synthetic data generation framework across multiple CVD datasets.

Dataset	Performance	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
SDDB [40]	Original Data	96.84	97.12	96.31	96.71	98.12
	Synthetic Data	92.45	93.14	91.74	92.42	95.04
MUSIC [41]	Original Data	97.35	97.85	96.91	97.39	98.47
	Synthetic Data	93.26	94.01	92.62	93.34	96.16
CVD Dataset [42]	Original Data	97.92	98.15	97.47	97.81	98.87
	Synthetic Data	93.89	94.71	92.95	93.82	96.47
Proposed Framework	Original Data	99.02	99.31	98.76	99.03	99.54
	Synthetic Data	95.14	96.38	94.05	95.20	97.34

4.1.10. Statistical Reliability and Significance Analysis

To evaluate the robustness of the suggested framework, the experiments were performed across 10 independent executions with different random initializations. Table 7 presents the classification performance obtained in each execution for Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The results indicate very consistent performance through all the executions with small deviation, which suggests that the suggested framework has a stable learning behavior and a reliable predictive capability rather than a performance that depends on a specific execution.

Table 8 presents the mean performance of the proposed framework with the corresponding standard deviations and the 95% confidence intervals over the 10 independent runs. The small confidence intervals and the small standard deviations indicate high statistical stability and low variability in the evaluation metrics. These results demonstrate that the reported classification performance is reliable and can be consistently reproduced under repeated experimental executions.

Table 7. Performance of the proposed framework over 10 independent experimental runs.

Run	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
1	94.98	96.24	93.89	95.05	97.21
2	95.11	96.33	94.02	95.17	97.3
3	95.29	96.55	94.18	95.34	97.43
4	95.07	96.29	94.01	95.14	97.28
5	95.18	96.44	94.11	95.25	97.36
6	95.1	96.36	94.04	95.18	97.34
7	95.24	96.48	94.15	95.3	97.39
8	95.02	96.27	93.94	95.09	97.25
9	95.2	96.46	94.12	95.27	97.38
10	95.21	96.38	94.06	95.22	97.46
Mean $\pm$ SD	95.14 $\pm$ 0.10	96.38 $\pm$ 0.11	94.05 $\pm$ 0.10	95.20 $\pm$ 0.10	97.34 $\pm$ 0.08

Table 8. Mean performance and 95% confidence intervals of the proposed framework.

Metric	Mean $\pm$ SD	95% Confidence Interval
Accuracy (%)	95.14 $\pm$ 0.10	95.07–95.21
Precision (%)	96.38 $\pm$ 0.11	96.30–96.46
Recall (%)	94.05 $\pm$ 0.10	93.98–94.12
F1-Score (%)	95.20 $\pm$ 0.10	95.13–95.27
ROC-AUC (%)	97.34 $\pm$ 0.08	97.28–97.40

Table 9 presents the statistical significance analysis of the proposed framework in comparison with the baseline methods. The reported *p*-values indicate whether the observed performance improvements are statistically significant. As all comparisons yield *p*-values below the conventional significance threshold of 0.05, the proposed framework demonstrates statistically significant improvements over the competing approaches. These findings confirm that the observed performance gains are unlikely to be attributed to random variation.

Table 9. Statistical significance analysis of the proposed framework compared with baseline methods.

Reference	Method	Accuracy Difference (%)	<i>p</i> -Value
Kita et al. [28] (2025)	DPMMs	6.4	<0.001
Lu et al. [29] (2023)	MTGAN-GRU	4.61	<0.001
Alqulaity et al. [31] (2024)	CGAN	3.46	<0.001
Kurakova et al. [32] (2025)	TVAE	2.67	<0.001
Jia et al. [33] (2025)	MSDGD	1.52	0.0004

4.1.11. Assessment of Feature Coverage and Statistical Consistency

Table 10 presents a comparative analysis of various GAN-based models and the proposed TabDDPM + LLaMA 2 + RAG framework in terms of total features supported, feature coverage, feature importance similarity, correlation preservation, and inference time per sample. To ensure fair comparison, we used these models in our pipeline to generate synthetic data from a real EHR dataset. The results demonstrate that GAN-based

approaches achieve lower inference time but exhibit moderate performance in terms of feature coverage and correlation preservation. In contrast, the proposed method achieves superior feature coverage (95%), higher similarity and correlation scores (0.93 and 0.92), and faster inference (18 ms/sample), highlighting its effectiveness and efficiency in generating high-quality synthetic EHR data.

Table 10. Performance comparison of synthetic EHR data generation models on feature fidelity and efficiency.

Reference	Model	Total Features	Feature Coverage (%)	Feature Importance Similarity	Correlation Preservation Score	Inference Time (ms/Sample)
Alqulaity et al. [31] (2024)	MedGAN	16	75%	0.70	0.68	22
Bozkurt et al. [43] (2026)	TableGAN	16	79%	0.74	0.71	24
Kumar et al. [44] (2023)	CGAN	16	80%	0.75	0.72	21
Parise et al. [45] (2025)	Conditional Tabular GAN	16	84%	0.81	0.78	27
Keskes et al. [46] (2025)	Conditional Table GAN	16	86%	0.83	0.81	29
Kang et al. [47] (2023)	CopulaGAN	16	82%	0.78	0.76	25
Pereira et al. [48] (2024)	Differentially Private GAN	16	77%	0.72	0.69	31
Kostopoulos et al. [49] (2025)	TVAE	16	83%	0.81	0.79	26
Proposed	TabDDPM + LLaMA 2 + RAG	16	95%	0.93	0.92	18

4.2. Ablation Study

Table 11 shows the ablation analysis evaluation of the contribution of each element in the proposed synthetic EHR data generation framework for cardiovascular risk prediction. In this analysis, only the investigated module was removed while all remaining model components, training settings, hyperparameters, and evaluation protocols were kept identical. The complete model, which uses the diffusion-based generation, LLM knowledge guidance, and RAG-based refinement, has great results in accuracy (95.14%) and minimum divergence values, indicating high predictive power and high statistical similarity to the original dataset. The elimination of the diffusion model leads to the highest measure of accuracy and increased divergence, making it essential for learning the underlying data distribution. Similarly, training steps such as quantile normalization and one-hot encoding lower the performance due to improper feature representation.

The lack of LLM knowledge guidance or RAG retrieval enhances distributional divergence, demonstrating the significance of clinical knowledge combination for identifying implausible patterns. In addition, constraint-based refinement reduction also slightly reduces accuracy and enhances statistical distance measures, proving its role in improving the realism and consistency of the generated synthetic data. Overall, the ablation experiment confirms that the combination of diffusion-based generation and knowledge-

based refinement is essential for generating high-quality and clinically plausible synthetic EHR data.

Table 11. Ablation study of the proposed synthetic EHR data generation framework.

Method	Accuracy (%)	KL Divergence	JS Divergence	Wasserstein Distance	Correlation Difference
without Diffusion Model	90.82	0.112	0.074	1.284	0.092
without Quantile Normalization	92.36	0.081	0.056	0.978	0.071
without One-Hot Encoding for Categorical Features	91.74	0.095	0.064	1.042	0.083
without LLM Knowledge Guidance	93.18	0.069	0.048	0.842	0.061
without RAG Knowledge Retrieval	93.74	0.063	0.045	0.789	0.054
without Constraint-Based Refinement	94.21	0.058	0.040	0.736	0.048
Proposed Model (Diffusion + LLM + RAG Knowledge Refinement)	95.14	0.055	0.037	0.686	0.043

4.3. Comparative Analysis

Table 12 presents the comparison of the proposed TabDDPM with the knowledge-guided refinement model with existing synthetic data generation methods. For a fair comparison, all methods were trained and evaluated using the identical preprocessing pipeline, training and testing partitions and evaluation metrics. The conventional methods such as DPMMs, MTGAN-GRU, CGAN and TVAE, show lower predictive accuracy and higher divergence scores, indicating limited ability to maintain the statistical characteristics of real EHR data. The diffusion-based methods enhance performance, as shown in the Gaussian diffusion model.

Table 12. Comparative performance analysis of synthetic EHR data generation framework.

Reference	Methods	Accuracy (%)	KL Divergence	JS Divergence	Wasserstein Distance	Correlation Difference
Kita et al. [28] (2025)	DPMMs	88.74	0.142	0.089	1.843	0.121
Lu et al. [29] (2023)	MTGAN-GRU	90.53	0.118	0.076	1.564	0.104
Alqulaity et al. [31] (2024)	CGAN	91.68	0.102	0.069	1.328	0.093
Kurakova et al. [32] (2025)	TVAE	92.47	0.089	0.058	1.112	0.081
Jia et al. [33] (2025)	MSDGD	93.62	0.071	0.047	0.902	0.064
Proposed	TabDDPM with Knowledge-Guided Refinement	95.14	0.055	0.037	0.686	0.043

Nevertheless, the proposed framework has the highest accuracy (95.14%) and the lowest KL, JS divergence, Wasserstein distance, and correlation difference. These findings show that integrating diffusion models with LLM and RAG-based knowledge refinement enhances the realism and statistical fidelity of generated synthetic EHR data.

4.4. Discussion

The proposed framework integrates TabDDPM-based synthetic EHR data generation with LLaMA 2 and RAG-based knowledge refinement to improve the clinical quality of the generated records. By combining distribution learning with knowledge-guided refinement, the framework generates synthetic EHR data that are statistically consistent and clinically meaningful for downstream cardiovascular risk prediction. The experimental results confirm the suitability of the proposed framework to achieve its goal of generating high-quality and statistically consistent synthetic EHR data for downstream prediction. The performance comparison reveals that the models trained on synthetic data achieve predictive accuracy close to that of models trained on the original dataset, demonstrating the high data utility. Besides statistical similarity measures, including KL divergence, JS divergence, Wasserstein distance, and correlation analysis ensure that the produced data maintains significant distributional characteristics and feature correlations of the actual dataset. The refinement process also enhances data quality, demonstrating the usefulness of integrating both the LLM-based reasoning and the diffusion-based data generation. The generalization capacity of the proposed framework is also observed through experiments that have been conducted on multiple datasets of CVD.

From the translational perspective, the proposed framework may help the development of clinical AI by providing a privacy-preserving alternative for generating representative healthcare datasets. In practical clinical settings, the generated synthetic EHR data can help to develop and validate cardiovascular risk prediction models when access to real patient records is restricted due to privacy regulations and ethical constraints. The framework may be able to be integrated with existing EHR analytics platforms or clinical decision support systems to help the development of AI-assisted risk assessment tools. However, prospective validation using real-world hospital data, evaluation across different patient populations and assessment within clinical workflows are required before clinical adoption. Other concerns, such as regulatory compliance, interpretability of model outputs, and clinician buy-in, must be addressed to achieve robust deployment in healthcare settings.

However, several limitations must be considered despite these promising results. The proposed framework involves a relatively high computational cost because the TabDDPM requires multiple forward and reverse diffusion steps during training and sampling, while the LLaMA 2 13B model combined with the RAG module introduces additional computational overhead during the knowledge-guided refinement stage. Consequently, the framework may require GPU acceleration and longer processing time when applied to large-scale EHR repositories. Although the framework is designed to be scalable, increasing dataset size and retrieval database complexity may increase training time, memory consumption, and retrieval latency necessitating more efficient indexing and distributed computing strategies for practical deployment. The current evaluation is primarily conducted using structured cardiovascular EHR benchmark datasets, which are relatively smaller and more homogeneous than real-world hospital databases. Large-scale clinical datasets collected from multiple healthcare institutions often contain heterogeneous patient populations, diverse coding standards, varying data quality and substantial missing or noisy information, which may influence the effectiveness of the proposed refinement strategy. Therefore, further validation on larger, multi-center and heterogeneous clinical datasets is required to comprehensively evaluate the robustness and generalizability of the pro-

posed framework. Future work will focus on improving computational efficiency through lightweight diffusion and LLM architectures incorporating domain-specific medical LLMs, optimizing the retrieval mechanism and extending the framework to multimodal healthcare data, including medical imaging, laboratory reports, and clinical notes, thereby enabling more scalable, privacy-preserving, and clinically applicable synthetic EHR data generation.

5. Conclusions

This study proposed a new framework for generating realistic synthetic EHR data to aid CVD risk prediction as well as for dealing with privacy and data accessibility issues. The model combines diffusion-based tabular data generation using TabDDPM with knowledge-guided refinement using LLaMA 2 13B with RAG. This combined method enhances the clinical plausibility and realism of the synthesized EHR data through the process of detecting and correcting implausible trends. The experimental findings show that the synthetic dataset retains key statistical characteristics and feature correlations of the original data while providing similar predictive performance in subsequent ML models. The proposed framework shows high generalization ability on several cardiovascular datasets. In conclusion, this work demonstrates that diffusion-based data generation with knowledge refining with the help of an LLM can generate high-quality and privacy-preserving synthetic healthcare data for reliable medical analytics and forecasting.

Author Contributions: All authors contributed to the conception of the problem setting and overall design of the work. K.B.A.S. built the conceptualization and methodology, J.S. implemented the work, and validation and writing were done by both. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data> (accessed on 9 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Corrêa, R.C.M.; Souza, J.M.d.; Nascimento, G.C.; Leonardo, P.T.; Oliveira, G.V.d. Exercise Selection and Rest Interval Duration Differentially Affect Post-Exercise Cardiac Autonomic Responses Following Resistance Training. *J. Vasc. Dis.* **2026**, *5*, 26. [CrossRef]
2. Tfelt-Hansen, J.; Garcia, R.; Albert, C.; Merino, J.; Krahn, A.; Marijon, E.; Basso, C.; Wilde, A.A.; Haugaa, K.H. Risk stratification of sudden cardiac death: A review. *Europace* **2023**, *25*, euad203. [CrossRef] [PubMed]
3. Holmstrom, L.; Chugh, H.; Nakamura, K.; Bhanji, Z.; Seifer, M.; Uy-Evanado, A.; Reinier, K.; Ouyang, D.; Chugh, S.S. An ECG-based artificial intelligence model for assessment of sudden cardiac death risk. *Commun. Med.* **2024**, *4*, 17. [CrossRef] [PubMed]
4. Kolk, M.Z.; Ruipérez-Campillo, S.; Wilde, A.A.; Knops, R.E.; Narayan, S.M.; Tjong, F.V. Prediction of sudden cardiac death using artificial intelligence: Current status and future directions. *Heart Rhythm.* **2025**, *22*, 756–766. [CrossRef] [PubMed]
5. Matjuda, E.N.; Nkeh-Chungag, B.N.; Engwa, G.A.; Sewani-Rusike, C.R.; Gubu-Ntaba, N.C.; Businge, C.B.; Saloñ, A.; De Boever, P.; Goswami, N. An Assessment of the Effect of HIV and ART on Cardiovascular Risk Factors to Predict Retinal Microvascular Impairment in Pregnant Women: A Pilot Study in a South African Population. *J. Vasc. Dis.* **2026**, *5*, 12. [CrossRef]
6. Marijon, E.; Narayanan, K.; Smith, K.; Barra, S.; Basso, C.; Blom, M.T.; Crotti, L.; d'Avila, A.; Deo, R.; Dumas, F.; et al. The Lancet Commission to reduce the global burden of sudden cardiac death: A call for multidisciplinary action. *Lancet* **2023**, *402*, 883–936. [CrossRef] [PubMed]
7. Liu, Y.; Acharya, U.R.; Tan, J.H. Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. *Comput. Methods Programs Biomed.* **2025**, *260*, 108571. [CrossRef] [PubMed]

8. Nicholas, I.; Kuo, H.; Gallego, B.; Jorm, L. Attention-based synthetic data generation for calibration-enhanced survival analysis: A case study for chronic kidney disease using electronic health records. *J. Biomed. Inform.* **2025**, *172*, 104928. [[CrossRef](#)]
9. Khan, S.A.; Murtaza, H.; Ahmed, M. Utility of GAN generated synthetic data for cardiovascular diseases mortality prediction: An experimental study. *Health Technol.* **2024**, *14*, 557–580. [[CrossRef](#)]
10. Goyal, M.; Mahmoud, Q.H. A systematic review of synthetic data generation techniques using generative AI. *Electronics* **2024**, *13*, 3509. [[CrossRef](#)]
11. Chatterjee, S.; Fruhling, A.; Kotiadis, K.; Gartner, D. Towards new frontiers of healthcare systems research using artificial intelligence and generative AI. *Health Syst.* **2024**, *13*, 263–273. [[CrossRef](#)]
12. Alkhalaf, M.; Yu, P.; Yin, M.; Deng, C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J. Biomed. Inform.* **2024**, *156*, 104662. [[CrossRef](#)] [[PubMed](#)]
13. Loni, M.; Poursalim, F.; Asadi, M.; Gharehbaghi, A. A review on generative AI models for synthetic medical text, time series, and longitudinal data. *npj Digit. Med.* **2025**, *8*, 281. [[CrossRef](#)] [[PubMed](#)]
14. Dalal, S.; Goel, P.; Onyema, E.M.; Alharbi, A.; Mahmoud, A.; Algarni, M.A.; Awal, H. Application of machine learning for cardiovascular disease risk prediction. *Comput. Intell. Neurosci.* **2023**, *2023*, 9418666. [[CrossRef](#)]
15. Cai, Y.; Cai, Y.Q.; Tang, L.Y.; Wang, Y.H.; Gong, M.; Jing, T.C.; Li, H.J.; Li-Ling, J.; Hu, W.; Yin, Z.; et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. *BMC Med.* **2024**, *22*, 56. [[CrossRef](#)] [[PubMed](#)]
16. Sumwiza, K.; Twizere, C.; Rushingabigwi, G.; Bakunzibake, P.; Bamurigire, P. Enhanced cardiovascular disease prediction model using random forest algorithm. *Inform. Med. Unlocked* **2023**, *41*, 101316. [[CrossRef](#)]
17. Yoon, J.; Mizrahi, M.; Ghalaty, N.F.; Jarvinen, T.; Ravi, A.S.; Brune, P.; Kong, F.; Anderson, D.; Lee, G.; Meir, A.; et al. EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digit. Med.* **2023**, *6*, 141. [[CrossRef](#)] [[PubMed](#)]
18. Li, J.; Cairns, B.J.; Li, J.; Zhu, T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digit. Med.* **2023**, *6*, 98. [[CrossRef](#)] [[PubMed](#)]
19. Rathnakumari, L.; Rao, G.R.K. Enhancing Heart Disease Prediction Through a CNN-GAN Hybrid Deep Learning Model. *SN Comput. Sci.* **2025**, *7*, 47. [[CrossRef](#)]
20. Zanella, L.; Zoppellaro, G.; Marigo, L.; Denas, G.; Padayattil, J.S.; Pengo, V. Risk factors for intracranial hemorrhage during vitamin K antagonist therapy in patients with nonvalvular atrial fibrillation: A case-control study. *Cardiovasc. Ther.* **2018**, *36*, e12458. [[CrossRef](#)] [[PubMed](#)]
21. Espitia, O.; Bruneval, P.; Assaraf, M.; Pouchot, J.; Liozon, E.; De Boysson, H.; Gaudric, J.; Chiche, L.; Achouh, P.; French Study Group for Large Vessel Vasculitides (GEFA); et al. Long-term outcome and prognosis of noninfectious thoracic aortitis. *J. Am. Coll. Cardiol.* **2023**, *82*, 1053–1064. [[CrossRef](#)] [[PubMed](#)]
22. Tsai, M.L.; Chen, K.F.; Chen, P.C. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *J. Am. Heart Assoc.* **2025**, *14*, e036946. [[CrossRef](#)] [[PubMed](#)]
23. Li, Y.; Mei, Z.; Sun, G.; Liu, Z.; Li, J.; Hock Ong, M.E.; Fan, H.; Cao, C. Machine learning for sudden cardiac death prediction among older adults using community-based electronic health records. *BMC Public Health* **2025**, *25*, 3521. [[CrossRef](#)] [[PubMed](#)]
24. Rujas, M.; del Moral Herranz, R.M.G.; Fico, G.; Merino-Barbancho, B. Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications. *Int. J. Med. Inform.* **2025**, *195*, 105763. [[CrossRef](#)] [[PubMed](#)]
25. Ibrahim, M.; Al Khalil, Y.; Amirrajab, S.; Sun, C.; Breeuwer, M.; Pluim, J.; Elen, B.; Ertaylan, G.; Dumontier, M. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Comput. Biol. Med.* **2025**, *189*, 109834. [[CrossRef](#)] [[PubMed](#)]
26. Osuala, R.; Kushibar, K.; Garrucho, L.; Linardos, A.; Szafranowska, Z.; Klein, S.; Glocker, B.; Diaz, O.; Lekadir, K. Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Med. Image Anal.* **2023**, *84*, 102704. [[CrossRef](#)] [[PubMed](#)]
27. Waseem, H.M.; Islam, S.U.; Matragkas, N.; Epiphaniou, G.; Arvanitis, T.N.; Maple, C. Review of generative AI for synthetic data generation: A healthcare perspective. *Artif. Intell. Rev.* **2026**, *59*, 55.
28. Kita, F.J.; Rao, G.S.; Kirigiti, P.J. Innovative synthetic EHR data generation: Diffusion models for enhanced privacy and clinical utility in multimorbidity clustering. *Comm. Sci.* **2025**, *37*, 2565163. [[CrossRef](#)]
29. Lu, C.; Reddy, C.K.; Wang, P.; Nie, D.; Ning, Y. Multi-label clinical time-series generation via conditional GAN. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 1728–1740.
30. Tian, M.; Chen, B.; Guo, A.; Jiang, S.; Zhang, A.R. Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2529–2539. [[CrossRef](#)] [[PubMed](#)]
31. Alqulaity, M.; Yang, P. Enhanced conditional GAN for high-quality synthetic tabular data generation in mobile-based cardiovascular healthcare. *Sensors* **2024**, *24*, 7673. [[CrossRef](#)] [[PubMed](#)]
32. Kurakova, A.; Homayouni, H. A comprehensive evaluation framework for synthetic medical tabular data generation. *J. Biomed. Inform.* **2025**, *171*, 104939. [[CrossRef](#)] [[PubMed](#)]

33. Jia, F.; Jia, F.; Li, H.; Qi, S.; Zhu, H. Breaking data barriers in medical diagnosis with MSDGD framework based on Gaussian Diffusion Generation. *Inf. Process. Manag.* **2025**, *62*, 104130. [CrossRef]
34. UCI Heart Disease Dataset. Available online: <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data> (accessed on 9 March 2026).
35. Yu, L.; Wu, J.; Wu, X.; Chen, C.; Chen, Y.; Fang, L.; Zheng, D. Interpretable machine learning model for cardiovascular disease risk prediction: A feature decomposition-based study. *BMC Public Health* **2025**, *25*, 3639. [CrossRef] [PubMed]
36. Najim, A.H.; Nasri, N. Artificial intelligence for heart disease prediction and imputation of missing data in cardiovascular datasets. *Cogent Eng.* **2024**, *11*, 2325635. [CrossRef]
37. Villaizán-Vallelado, M.; Salvatori, M.; Segura, C.; Arapakis, I. Diffusion models for tabular data imputation and synthetic data generation. *ACM Trans. Knowl. Discov. Data* **2025**, *19*, 1–32. [CrossRef]
38. Liang, Y.L.; Chen, C.F.; Wu, M.H.; Hsu, K.F.; Chang, P.T.; Chuang, Y.J.; Chen, P.F. Fine-tuning Llama-2-13B with AI-generated medical diagnoses: A novel strategy for optimizing ICD coding in gynecologic oncology. *Taiwan J. Obstet. Gynecol.* **2025**, *64*, 978–984. [CrossRef] [PubMed]
39. Kresevic, S.; Giuffrè, M.; Ajcevic, M.; Accardo, A.; Crocè, L.S.; Shung, D.L. Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. *npj Digit. Med.* **2024**, *7*, 102. [CrossRef] [PubMed]
40. Sudden Cardiac Death Database. Available online: <https://physionet.org/content/music-sudden-cardiac-death/1.0.1/> (accessed on 9 March 2026).
41. MUSIC Dataset. Available online: <https://physionet.org/content/sddb/1.0.0/> (accessed on 9 March 2026).
42. CVD Dataset. Available online: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset> (accessed on 9 March 2026).
43. Bozkurt, M.H. Bridging stability and utility in synthetic tabular data: The sTableGAN model. *Knowl. Based Syst.* **2026**, *336*, 115303. [CrossRef]
44. Kumar, V.; Sinha, D. Synthetic attack data generation model applying generative adversarial network for intrusion detection. *Comput. Secur.* **2023**, *125*, 103054. [CrossRef]
45. Parise, O.; Kronenberger, R.; Parise, G.; de Asmundis, C.; Gelsomino, S.; La Meir, M. CTGAN-driven synthetic data generation: A multidisciplinary, expert-guided approach (TIMA). *Comput. Methods Programs Biomed.* **2025**, *259*, 108523. [CrossRef] [PubMed]
46. Keskes, M.O.H.A.M.E.D. Generative adversarial networks for synthetic data generation in deep learning applications. *J. Artif. Intell. Res. Innov.* **2025**, *1*, 028–033. [CrossRef]
47. Kang, H.Y.J.; Batbaatar, E.; Choi, D.W.; Choi, K.S.; Ko, M.; Ryu, K.S. Synthetic tabular data based on generative adversarial networks in health care: Generation and validation using the divide-and-conquer strategy. *JMIR Med. Inform.* **2023**, *11*, e47859. [CrossRef] [PubMed]
48. Pereira, M.; Kshirsagar, M.; Mukherjee, S.; Dodhia, R.; Lavista Ferres, J.; De Sousa, R. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLoS ONE* **2024**, *19*, e0297271. [CrossRef] [PubMed]
49. Kostopoulos, G.; Fazakis, N.; Kotsiantis, S.; Dimakopoulos, Y. Enhancing Semi-Supervised Learning in Educational Data Mining Through Synthetic Data Generation Using Tabular Variational Autoencoder. *Algorithms* **2025**, *18*, 663. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.