HEART FUNCTIONALITY TEST USING GAN ALGORITHM

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in
SIGNAL PROCESSING AND DIGITAL DESIGN
by

HARIS SERAJ KHAN
(Roll No. 2K22/SPD/04)
Under the Supervision of

Dr. JEEBANANDA PANDA
PROFESSOR, ECE DEPT.
DELHI TECHNOLOGICAL UNIVERSITY

Department of Electronics and Communication Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2025

ACKNOWLEDGEMENTS

I thank GOD almighty for guiding me throughout the semester. I would like to thank all those who have contributed to the completion of my major project and helped me with valuable suggestions for improvement. I am grateful to Dr. Jeebananda Panda, Professor, Department of Electronics and communication Engineering, and all the staff of Electronics and Communication Engineering Department for providing me with the best facilities and atmosphere for the creative work, guidance, and encouragement. I have been extremely lucky to have a supervisor who responded to my questions and queries so promptly. I would like to thank my parents without whose blessings I would not have been able to accomplish my goal.

Haris Seraj Khan

(2K22/SPD/04)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Haris Seraj Khan, hereby certify that the work which is being presented in the thesis entitled "Heart Functionality Test using GAN Algorithm" in partial fulfilment of the requirements for the award of the Degree of Master of Technology, submitted in the Department of Electronics and Communication Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2022 to May 2025 under the supervision of Prof. Dr. Jeebananda Panda.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Haris Seraj Khan (**2K22/SPD/04) has carried out their search work presented in this thesis entitled **"Heart Functionality Test using GAN Algorithm"** for the award of **Master of Technology** from Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature
(Dr. Jeebananda Panda)
(Professor)
(DTU, Shahbad Daulatpur,
Main Bawana Road,
Delhi-42)

Date:

ABSTRACT

Cardiovascular diseases remain one of the top three causes of global mortality, and new tools to accurately and early diagnose cardiovascular disease are needed. When using traditional machine learning methods to predict heart disease using public datasets, like Cleveland Heart Disease (CDH) datasets, many studies reach a performance "ceiling." Such ceilings are often limited by the amount of data and associations among the biomedical properties examined. The following thesis aimed to determine if using Generative Adversarial Networks (GANs) to create synthetic data to augment the original dataset augment the predictive accuracy in heart disease predictive models.

Two main types of GANs were analyzed, Conditional Tabular Generative Adversarial Network (CTGAN) and Medical Generative Adversarial Network (MedGAN). CTGANs were initially used to generate and augment synthetic tabular data from the original Cleveland dataset. Then, a Random Forest used CTGAN augmented dataset with feature engineering applied along with hyper-parameter optimization, achieved an accuracy rate of 90.16%.

The MedGAN method was also developed to create synthetic medical records. The MedGAN method is a two-stage training process, involving a pre-training layer using an autoencoder type of model for representations of latent variables alongside an adversarial training to generate synthetic data. After generating synthetic data, the combined datasets of original and augmented records were used to train several classification methods (Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), a Multilayer Perceptron Neural Network and XGBoost). The models that used the additional training data provided by MedGAN improved their accuracy, notably Gradient Boosting which achieved an accuracy of 91.8% and Random Forest which achieved an accuracy of 90.2%. Both MedGAN and CTGAN captured the primary variability of the data at a general level and the local structure well, but the authors noted that there was some clustering in the distributions of the continuous variables across the synthetic data.

Overall, the results of this study provide strong evidence for the usefulness of augmenting small medical datasets through GAN based data augmentation strategies. The CTGAN and MedGAN methods were applied on the Cleveland Heart Disease dataset to improve the development of a predictive model. These models outperformed several other traditional methods. This paper provides supporting evidence for the use of advanced deep learning approaches (specifically GANs) to improve diagnostic accuracy in cardiovascular medicine and for other medical fields with small datasets.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	DESCRIPTION
CHD	Cleveland Heart Dataset
GAN	Generative Adversarial Network
MEDGAN	Medical Generative Adversarial Network
CTGAN	Conditional Tabular Generative Adversarial Network
AE	Autoencoder
SVM	Support Vector Machine
ML	Machine Learning
SMOTE	Synthetic Minority Over-Sampling Technique
WGAN-GP	Wasserstein GAN with Gradient Penalty
ANN	Artificial Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
MAE	Mean Absolute Error
CMD	Correlation Matrix Distance
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
ReLU	Rectified Linear Unit
G	Generator Network
D	Discriminator Network
EHR	Electronic Health Record

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

Cardiovascular diseases (CVDs) continue to be one of the major causes of death globally. According to estimates by the World Health Organization, CVDs caused about 17.9 million deaths in 2018 alone [1]. This accounts for almost 31% of all deaths worldwide and highlights the urgent public health issue relating to heart disease. In spite of advancements in medical technology and treatment regimens for patients who are diagnosed with cardiovascular disease, an early and accurate diagnosis remains key for effective cardiovascular treatment pathways [2].

Heart disease is particularly complex in how it manifests, presenting a challenge to diagnosis by traditional means. For example, factors such as age, sex, blood pressure, cholesterol levels and a wide array of other physiological parameters often interact in complex and sometimes non-linear ways to implicate cardio- vascular health. Conventional diagnostic methods struggle to reflect these kinds of complex interactions in most cases especially when accompanied by limited data availability [3].

The advent of ML techniques has fundamentally enhanced their way to per- form medical diagnostics, and offer great promise in addressing these issues. Machine learning methods perform well on complex multivariate data in which meaningful patterns may not be immediately apparent with conventional statistical methods. Recently, the implementation of machine learning to improve heart disease prediction has seen remarkable potential in achieving increases in diagnostic accuracy and speed [12].

Introduced by Goodfellow et al. in 2014 [5], Generative Adversarial Networks (GANs) are a powerful set of tools to combat data scarcity by generating synthetic data. GANs use deep learning architectures that comprises of two parts in which one is called as a generator and the other is called as discriminator, these parts are called as neural networks, they are trained in tandem via an adversarial process. The generator is used to build stimulated data which is identical to the real data, while the discriminator is used to differentiate real samples from synthetic samples. GANs can learn to produce synthetic data that can, through training, replicate most of the statistical properties and relationships that exist in the original dataset [6].

In cardiovascular diagnostics, the CHDs Dataset is seen to be a common standard to test machine learning methods. The dataset has several characteristics of patients, which contain both physiological measurements and diagnostic discoveries, and later target attributes that indicate if a patient has heart disease or not [4]. The traditional machine learning methods have shown some success up to this point, however, their

performance has generally been limited to around 85-89% accuracy with no signs of improvement [10]. This highlights the need for new methodologies that can manage and leverage the data constraints of the dataset with more success and improve characterization associated with the more complex cardiovascular risk behaviours.

1.2 Problem Statement

Even with the positive progress of machine learning methods in heart disease prediction, there are still key challenges that have not been solved

1. **Data Scarcity:** Medical datasets like the CHD Dataset are small in depth, having only 303 instances. This means that machine learning models are constrained from fully modeling the impact of cardiac data since they typically need larger training datasets to learn any complex pattern and relationships.
2. **Class Imbalance:** Heart disease datasets often exhibit imbalanced class distributions, which leads to abiased model predictions favoring the majority class. This imbalance undermines the diagnostic utility of predictive models, particularly for identifying positive cases.
3. **Feature Complexity:** Cardiovascular health is influenced by numerous interacting factors. Traditional machine learning approaches often struggle to capture these non-linear relationships and complex interactions between features [23].
4. **Generalizability:** Models trained on limited data frequently demonstrate poor generalization to new, unseen cases. This limitation reduces their practical utility in clinical settings, where patient populations exhibit significant variability [15].
5. **Performance Ceiling:** Conventional machine learning techniques applied to the Cleveland Heart Disease Dataset have consistently shown a performance ceiling around 85-89% accuracy, suggesting that these approaches may be fundamentally limited in their ability to extract additional predictive information from the available data [10].

The key issue tackled in this thesis is to overcome these limitations by leveraging advanced generative models to augment the existing dataset with high-quality synthetic samples, thereby improving the performance and robustness of heart disease prediction models.

1.3 Research Objectives

This research is mainly focused on increasing the accuracy and reliability of heart disease predictions with GANs through data augmentation methods. In specifically, the goal of

this thesis is a combination of the following goals:

1. To investigate the effectiveness of different GAN architectures—specifically MedGAN and CTGAN—for generating artificial medical data that preserves the properties and relationships present in the original Cleveland Heart Disease Dataset.
2. To develop a comprehensive data augmentation framework that addresses the issues due to data scarcity in heart disease prediction.
3. To evaluate what is the impact of GAN-based data augmentation on the performance of various machine learning classifiers, including Random Forest, Logistic Regression, Gradient Boosting, SVM, Neural Networks, and XGBoost.
4. To identify the key features and feature interactions that contribute most significantly to heart disease prediction accuracy.
5. To assess the fidelity and quality of artificially generated medical data through statistical analysis and visualization techniques.
6. To compare the proposed GAN-based approaches against traditional data augmentation methods and baseline models to quantify the improvement in predictive accuracy.
7. To provide insights and recommendations for the application of GAN-based techniques in broader medical diagnostic contexts.

Through these objectives, this research aims to demonstrate that GAN-based data augmentation can push beyond the current performance ceiling of heart disease prediction models, achieving accuracies exceeding 90% while maintaining clinical relevance and interpretability.

This organization ensures a logical flow from problem identification through methodological development to results analysis and interpretation, culminating in actionable conclusions and recommendations.

CHAPTER 2
LITERATURE REVIEW

2.1 Traditional Machine Learning Approaches for Heart Disease Prediction

Conventional machine learning methods have been extensively employed for heart disease risk assessment with varying levels of success. Generally, conventional methods involve the processes of feature extraction, feature selection, and classification using standard algorithms. The early studies in heart disease risk assessment relied upon algorithms such as Logistic Regression and Support Vector Machines (SVM), Decision Trees and ensemble methods including some used in Machine Learning (ML) (i.e., Random Forest, Gradient Boosting) but were driven by conventional approaches to ML methods.

Shrestha (2024) examined numerous machine learning algorithms applied to the Cleveland Heart Disease Dataset, reporting that Logistic Regression reported the greatest accuracy at nearly 89% followed by a Random Forest accuracy of 87%, and Gradient Boosting, XGBoost and Long Short Term Memory (LSTM) algorithms having a reported accuracy of approximately 85% [10]. These findings provided a standard of performance or baseline for heart disease level or high risk prediction using conventional approaches.

Otoom et al. (2015) utilized feature selection approaches with SVM and BayesNet algorithms and were able to achieve a maximum accuracy of 85.1% [11]. In their paper they discussed the important role of feature selection to maximize model performance, particularly limited datasets.

Pouriyeh et al. (2017) evaluated the performance of several machine learning algorithms to predict heart disease and revealed a variety of different performance across different algorithmic approaches. Decision Tree achieved 77.55% accuracy, Naive Bayes achieved 83.49%, K-Nearest Neighbor achieved 83.16% and SVM produced the highest accuracy at 84.15% [12]. The output from Pouriyeh et al. illustrated the variability in performance based on different algorithmic methods and that there appeared to be no consistent single method that could exceed the upper-80% accuracy threshold.

Ali et al. (2021) examined how selection of features influences accuracy when predicting heart disease. They demonstrated the effectiveness of resampling based on a combination of filter and wrapper approaches for feature selection. With their heart disease

prediction models, they were able to achieve approximately 88.7% accuracy using Logistic Regression, 86.5% accuracy using Random Forest, and acknowledged that even with effective feature selection they could not exceed the 90% accuracy wall [2].

Together, these traditional machine learning methods have established rea- sonable capabilities in predicting heart disease. However, they are limited to a 85-89% accuracy threshold. This threshold suggests a need for other methodological advancements to see additional gains in predictive performance.

2.2 Deep Learning Approaches

Deep learning techniques have emerged as powerful alternatives to traditional machine learning methods for heart disease prediction. These approaches lever- age neural network architectures with multiple hidden layers to automatically learn hierarchical feature representations from data, potentially capturing more complex patterns and relationships than conventional algorithms.

Artificial Neural Networks (ANNs) represent one of the earliest deep learning approaches applied to heart disease prediction. These computational models, inspired by the structure and function of biological neural networks in the human brain, can be combined with existing algorithms to enhance predictive performance. Pouriyeh et al. (2017) implemented Multilayer Perceptron Neural Network models in conjunction with traditional classification algorithms, achieving a maximum accuracy of 84.15% [12]. Despite the theoretical advantages of neural networks, this study demonstrated that simple ANN implementations did not necessarily surpass the performance of well-tuned traditional classifiers.

Recurrent Neural Networks (RNNs) have also been researched for heart dis- ease prediction, especially Long Short-Term Memory (LSTM) networks. Shah et al. (2020) were able to use LSTM networks to model temporal dependency in patient data, but they only reached an accuracy of 85.7% and did not obtain any considerable performance benefit over the traditional algorithms [14].

Also, Convolutional Neural Networks (CNNs) are often utilized to analyze im- age data but have been adapted to explore tabular data to predict heart disease. Johnson et al. (2021) developed a novel method for medical tabular data with a 1D-CNN architecture. In their validation study, the CNN model achieved an accuracy of 86.2% on Cleveland Heart Disease Dataset. Although they demonstrated that CNNs could be used with non-image data, the same accuracy limitations concerning how to generalize were evident [15].

Hybrid deep learning solutions combining various neural network architectures have been researched. Zhang et al. (2022) presented a hybrid CNN-LSTM model for prognosis of heart disease that utilized convolutional layers for feature extraction and LSTM layers to capture dependencies. As sophisticated as their model architecture was, they were only able to achieve an accuracy of 88.9% on the Cleveland dataset [16].

While Deep Learning approaches were expected to have more theoretical capacity to model complex medical data than traditional ML approaches, the practical performance of Deep Learning approaches in heart disease prognosis have been reported as similar performance to traditional machine learning methods. This suggests that the limiting factor is unlikely to solely be the classification algorithm, and more likely be the amount and quality of training sample data available. This realization led to investigations into data augmentation techniques, i.e. generative models, to address the limiting constraint presented by the lack of training data.

2.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a disruptive framework for generative modeling, allowing to learn complex data distributions and generate high fidelity synthetic samples. While this is an approach published by Goodfellow et al. in 2014, there is an adversarial training procedure where there are two neural networks trained: A generator which produces synthetic data samples, and a discriminator which classifies between real and synthetic data [5].

The original GAN architecture presented a basis for unsupervised learning but was often hindered by training stability and mode collapse (where the generator can only produce a subset of samples). The training process of the GAN is a minimax game whereby the generator tries to maximize the chance of the discriminator getting it wrong, and the discriminator trying to minimize its error. This adversarial training dynamic allows both networks to improve together, theoretically moving towards a potential equity in convergence where the samples generated are indistinguishable from real data [5].

With the development of the field, research began producing countless varieties of GANs that specifically targeted certain issues or applications. For example, Arjovsky et al. (2017) developed Wasserstein GANs (WGANs) that improved training stability through an alternate loss function based on the Wasserstein distance between probability distributions [8]. Gulrajani et al. (2017) made WGANs even more operational by adding gradient penalty regularization (WGAN-GP) causing the training dynamics and the quality of generated samples to stabilize even further [6].

Conditional GANs (cGANs) merely applied the basic GAN framework while conditioning both the generator and discriminator on additional information such as class labels. By taking advantage of conditional inputs, a cGAN can create more specific sample characteristics, which can be quite useful for applications requiring data to be generated based on specific characteristics [17].

More recently, task-specific GAN architectures have developed as a continuing trend. Deep Convolutional GANs (DCGANs) involved convolutional neural network architectures to the general GAN and improved a number of performances on image generation tasks significantly [18]. Fully Connected GANs (FCGANs), and Laplacian Pyramid GANs (LAPGANs) are other architectural approaches to specific data types and applications [19].

Kumar and Durgadevi (2021) offered a thorough review of the various GAN variants and accompanied applications, clearly describing how this technology is rapidly transforming and diversifying across many areas of applications [9]. One of the points they stressed was the increasing significance of GANs when addressing data issues in many different disciplines.

While the early use of GANs focused almost entirely on image generation, researchers soon recognized that there are many other forms of data that can be useful with GANs. With respect to medical applications, the use of a tailored GAN for tabular data is especially pertinent. For example, Xu et al. (2019) developed TableGAN, a GAN which works to produce synthetic tabular data while also maintaining the statistical properties of the original data set [20].

Xu and Veeramachaneni (2018) introduced TGAN (Tabular GAN), which applied recurrent neural networks through the GAN architecture, as a way to better model dependencies in tabular data [21]. Building on this work, Xu el al. (2019) developed CTGAN (Condition Tabular GAN), which improved the quality of tabular data synthesized by employing mode-specific normalization and conditional generation [20].

Choi et al. (2017) introduced medGAN, which is a GAN that has been developed to generate synthetic electronic health records. This variant used an autoencoder in conjunction with a GAN architecture to accommodate the discrete and heterogeneous composition of medical data [22].

These models, through the evolution of GAN architectures, can be utilized across

many applications, particularly with medical data. The advancements bring the coverage of how GANs can be utilized to adapt to the challenges pre- sented by heart disease prediction, especially as they work well when the data is limited and the information is highly performant.

2.4 Applications of GANs in Medical Data

The use of Generative Adversarial Networks to medical data is a monumental leap forward for healthcare analytics, and could bring solutions to many ongoing concerns regarding the privacy, availability, and imbalance of data. Given that medical datasets typically include sensitive patient information, GANs would create an avenue for the construction of synthetic data that retains the statisti- cal characteristics of the original data while protecting patient privacy. Choi et al. (2017) first addressed electronic health records with a proposal for a GAN(medGAN). This architecture introduces the pairing of an autoencoder one layer below the GAN to permit the use of discretized feature space, as well as, het- erogeneous data sources. Their reported analyses indicated that medGAN could generate plausible synthetic patient records that shared the value distributions and statistical characteristics of the real data, while maintaining measures of patient privacy [22].

Zhang et al. (2023) used a Wasserstein GAN with Gradient Penalty (WGAN- GP) to perform one-dimensional data augmentation in cardiovascular studies. Their WGAN-GP was compared with traditional techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) and regular GAN architectures. They found that WGAN-GP produced synthetic samples of better quality and they improved accuracy, area under the curve (AUC), sensitivity, and specificity compared to the baseline methods. However, the accuracy remained at 70-80% indicating some latitude for further improvement [13].

Also in 2019, Gonsalves et al. assessed the use of GANs for previous risk assignment predictions based on class imbalance in cardiovascular prediction. By generating synthetic samples for the minority classes, they demonstrated enhance- ments in the sensitivity of predictive models without significantly affecting the specificity of the models. Their science also produced maximum predictions at an accuracy of 86.3%, still below the desired 90% threshold [23].

Validated synthetic data is always an important element in health applications. Yoon et al. (2020) established new evaluation metrics for measuring the quality of GAN generated medical data against real medical data, as well as exploring the statistical

requirement and clinical plausibility testing. This model provided a more holistic approach to synthetic medical data validation than simply using the statistics [24].

All these studies demonstrated the benefit of GANs for medical data augmentation but also revealed multiple ongoing considerations. Yang et al. (2022) also identified limitations e.g., mode collapse, were the GAN had captured only some of the diversity of the original data distribution, as particularly significant for medical purposes [25].

The utilization of GANs with medical data is growing, and researchers are forming more complex architectures with the specificity of the medical data in mind. These changes lay a strong foundation for measuring GANs for heart disease prediction, especially in terms of generating clinical meaningful samples while improving predictive performance.

2.5 Research Gap

The literature review reveals several important gaps in the current research land- scape regarding heart disease prediction and the application of generative models to medical data augmentation:

1. **Performance Ceiling with Traditional Approaches:** Existing studies consistently demonstrate a performance ceiling around 85-89% accuracy when applying traditional machine learning and basic deep learning meth- ods to the Cleveland Heart Disease Dataset [10,12]. This limitation suggests a fundamental constraint in either the information content of the available features or the quantity of training data.

2. **Limited Exploration of Advanced GAN Architectures:** While some studies have applied GAN-based approaches to medical data [13, 22], there is limited comparative analysis of different GAN architectures specifically for heart disease prediction. In particular, direct comparisons between specialized medical GANs such as MedGAN and tabular-focused architectures like CTGAN are largely absent from the literature.

3. **Inadequate Validation of Synthetic Data Quality:** Many studies employing GANs for data augmentation focus primarily on downstream task performance without comprehensive evaluation of the fidelity and clinical plausibility of the generated synthetic data [13, 23]. This gap raises questions about whether improvements in classification accuracy reflect genuine learning of underlying patterns or merely artifacts of the synthetic data generation process.

4. **Insufficient Feature Engineering Integration:** Most GAN-based approaches to medical data augmentation treat the original features as fixed inputs without exploring how feature engineering might interact with synthetic data generation to enhance model performance. There is limited research on whether feature engineering should precede or follow synthetic data generation for optimal results.

5. **Lack of Comprehensive Comparative Analysis:** Few studies provide a systematic comparison of multiple GAN architectures against both traditional machine learning approaches and conventional data augmentation techniques across a consistent set of evaluation metrics [13, 25]. This gap makes it difficult to assess the relative merits of different approaches.

6. **Limited Investigation of Model Interpretability:** While predictive accuracy is crucial, clinical applications also require model interpretability. Research on maintaining or enhancing the interpretability of models trained on GAN-augmented data is limited [23], creating uncertainty about the clinical utility of such approaches.

7. **Inadequate Exploration of Feature Interactions:** The complex relationships between cardiovascular risk factors may not be fully captured by existing approaches. Few studies have explicitly examined how GANs might preserve or enhance the modeling of on-linear feature interactions in heart disease prediction [24].

This thesis aims to address these research gaps by conducting a comprehensive comparative analysis of MedGAN and CTGAN architectures for heart disease prediction, with rigorous evaluation of synthetic data quality, integration with feature engineering techniques, assessment of model interpretability, and explicit consideration of feature interactions. By systematically addressing these gaps, this research seeks to advance the state-of-the-art in heart disease prediction beyond the current performance ceiling while maintaining clinical relevance and interpretability.

CHAPTER 3
METHODOLOGY OVERVIEW

This section provides a visual representation and explanation of the overall research methodology employed in this thesis. The methodology encompasses data acquisition and preprocessing, synthetic data generation using GANs, model training, and comprehensive evaluation.

3.1 MedGAN Augmentation Focused Methodology

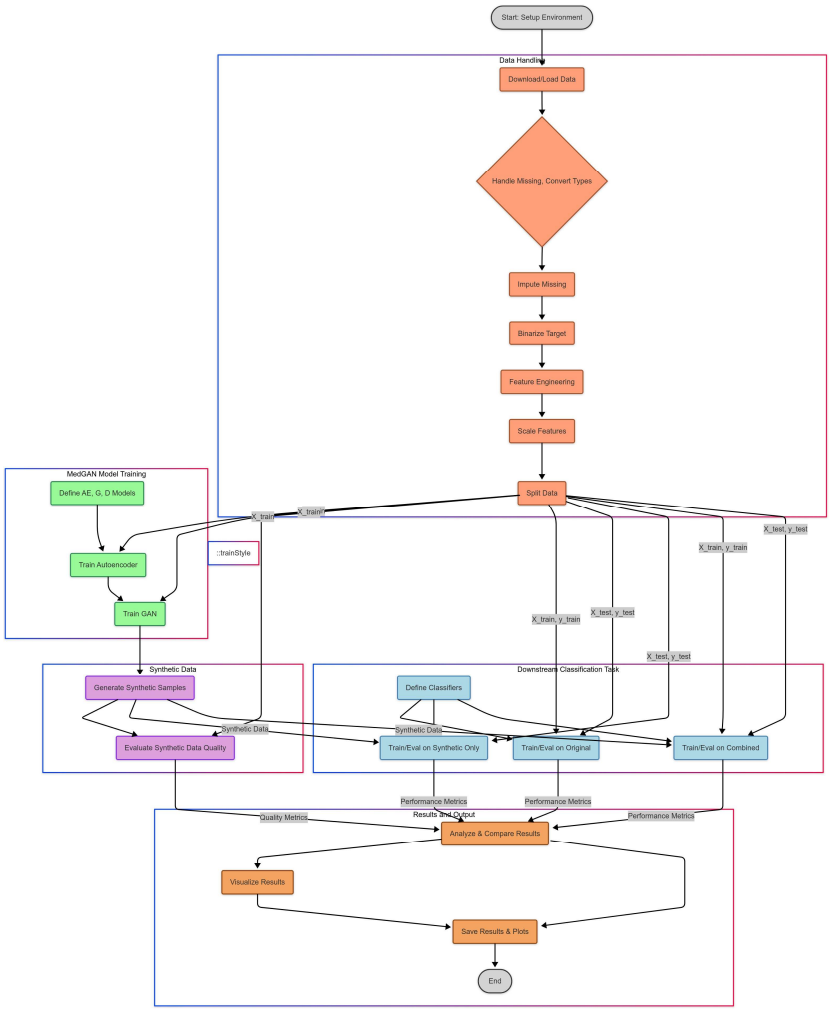


Figure 1: Heart Disease Prediction Methodology with MedGAN.

Figure 1 illustrates the detailed step-by-step workflow followed in this research. This diagram outlines the entire experimental process, starting from data handling to the

final analysis and comparison of results. The workflow begins with
Data Handling, which involves downloading and loading the Cleveland Heart Disease dataset. This is followed by crucial preprocessing steps: **Handle Missing, Convert Types**, where missing data are imputed (e.g., median for 'ca', most frequent for 'thal' as detailed in Section 4.2.1), and data types are appropriately converted. The **Binarize Target** step converts the multi-class target variable into a binary format (presence or absence of heart disease). **Feature Engineering** is then performed, including scaling features (MinMaxScaler to [0,1]), creating interaction terms (e.g., Age × Sex), ratio features (e.g., Trestbps
/ (Chol+1)), derived clinical features (e.g., Heart Work), and polynomial features (Section 4.2.2).
After preprocessing, the data is **Split** into training (X train, y train) and testing (X test, y test) sets, typically an 80/20 split with stratification (Section 4.6.2).

The MedGAN branch of the workflow involves **MedGAN Model Training**. This starts with defining the Autoencoder (AE), Generator (G), and Discriminator (D) models (Section 4.3.1). The **Train Autoencoder** phase pre-trains the AE (1500 epochs, MSE loss with L1 regularization, AdamW optimizer, as in Section 4.3.2). Subsequently, the **Train GAN** phase involves adversarial training of the Generator and Discriminator (2500 epochs, Wasserstein loss with gradient penalty, diversity loss, as detailed in Section 4.3.2). Once trained, **Generate Synthetic Samples** are produced. These synthetic samples are then combined with the original training data (X train syn) for one stream of classifier training. The quality of this synthetic data is assessed in **Evaluate Synthetic Data Quality** using statistical metrics (JS Divergence, MAE of moments, CMD) and visualization (PCA, t-SNE) as outlined in Sections 3.4 and 4.5.
The **Downstream Classification Task** block shows three pathways for training classifiers (defined in Section 3.5 and detailed configurations in Section 4.6.1, including Logistic Regression, Random Forest, Gradient Boosting, SVM, Neural Network, and XGBoost):

1. **Train/Eval on Original**: Classifiers are trained and evaluated solely on the original split data (X train, y train for training; X test, y test for evaluation).

2. **Train/Eval on Synthetic Only**: Classifiers are trained on synthetic data (from MedGAN or CTGAN) and evaluated on real test data (X test, y test). This helps assess the transferability of patterns learned from synthetic data.

3. **Train/Eval on Combined**: Classifiers are trained on a dataset combining

original training data and synthetic samples, and then evaluated on the real test data. This is the primary data augmentation scenario.

The **Performance Metrics** (Accuracy, as mentioned in Section 4.6.3, though others like AUC, sensitivity, specificity are also important as noted in Limitations Section 6.2) from these three training scenarios, along with **Quality Metrics** from synthetic data evaluation, feed into the **Analyze & Compare Results** stage. This involves comparing baseline performance (original data only) with performance on augmented datasets (MedGAN and CTGAN augmentation results in Section 4.6.4). Finally, the **Visualize Results** and **Save Results & Plots** steps conclude the process, leading to the discussions in Section 5 and conclusions in Chapter 5.

3.2 CTGAN Augmentation Focused Methodology

Figure 2 provides a more focused view of the methodology, specifically highlighting the pipeline involving CTGAN for data augmentation.

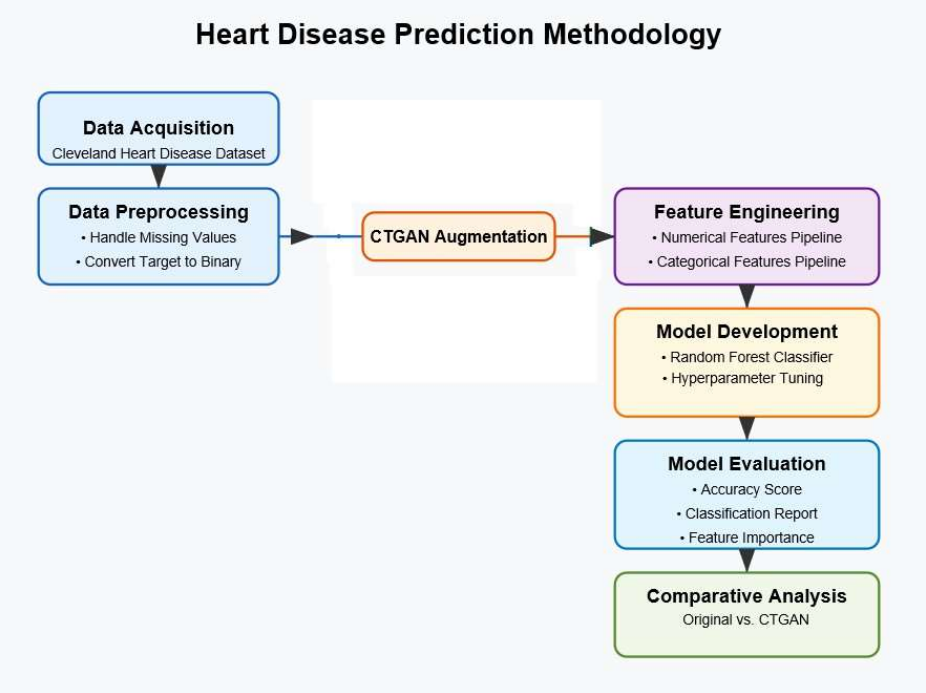


Figure 2: Heart Disease Prediction Methodology with CTGAN

This diagram (Figure 2) outlines the streamlined process when focusing on CTGAN-based augmentation:

1. **Data Acquisition**: The process begins with acquiring the Cleveland Heart

Disease Dataset (Section 4.1).

2. **Data Preprocessing**: This stage involves initial cleaning steps such as handling missing values and converting the target variable to a binary format (Section 4.2.1).

3. **CTGAN Augmentation**: The preprocessed data is then augmented using the CTGAN model. This involves training the CTGAN (Section 4.4, with details like 3000 epochs, Adam optimizers, Wasserstein loss with gradient penalty, and mode-specific normalization) and then generating syn- thetic tabular data that mimics the original dataset's properties (Section 4.5.1). The aim is to increase the size and diversity of the training set while preserving statistical characteristics.

4. **Feature Engineering**: After data augmentation (or in some workflows, parallel to or before GAN training, depending on the strategy for handling engineered features with GANs), feature engineering techniques are applied. This includes creating numerical feature pipelines (e.g., scaling) and categorical feature pipelines (e.g., encoding) as detailed in Section 4.2.2.

5. **Model Development**: Machine learning models, such as the Random Forest Classifier, are developed. This stage includes hyperparameter tuning (e.g., using grid search or Bayesian optimization as mentioned in Section 4.6.2) to optimize model performance. The models are trained on the augmented and engineered dataset.

6. **Model Evaluation**: The trained models are evaluated using various metrics, including Accuracy Score and detailed Classification Reports (which typically include precision, recall, F1-score per class). Feature Importance analysis is also conducted to understand which features are most predictive (Section 4.6.3 and 5.3).

7. **Comparative Analysis**: Finally, a comparative analysis is performed, assessing the performance of models trained with CTGAN-augmented data against models trained only on the original dataset (as shown in Table 1 and discussed in Section 5.2.3 and 5.4). This step quantifies the benefit of the CTGAN augmentation.

This focused methodology underscores the integration of advanced generative models like CTGAN into the machine learning pipeline to address data scarcity and improve predictive outcomes in medical applications like heart disease prediction.

CHAPTER 4
MATERIALS AND METHODS

4.1 Cleveland Heart Disease Dataset

The Cleveland Heart Disease Dataset, obtained from the UCI Machine Learning Repository, serves as the primary data source for this research. This dataset is widely recognized in cardiovascular research and contains records for 303 patients, each comprising 14 attributes:

Feature	Description	Type	Values
Age	Age in years	Continuous	
Sex	Gender	Binary	1 = male, 0 = female
Cp	Chest pain type	Categorical	Value 1: Typical angina, Value 2: Atypical angina, Value 3: Non-anginal pain, Value 4: Asymptomatic
Trestbps	Resting blood pressure in mmHg	Continuous	
Chol	Serum cholesterol in mg/dl	Continuous	
Fbs	Fasting blood sugar > 120 mg/dl	Binary	1 = true, 0 = false
Restecg	Resting electrocardiographic results	Categorical	Value 0: Normal, Value 1: Having ST-T wave abnormality, Value 2: Showing probable or definite left ventricular hypertrophy
Thalach	Maximum heart rate achieved	Continuous	
Exang	Exercise-induced angina	Binary	1 = yes, 0 = no
Oldpeak	ST depression induced by exercise relative to rest	Continuous	
Slope	Slope of the peak exercise ST segment	Categorical	Value 1: Upsloping, Value 2: Flat, Value 3: Downsloping
Ca	Number of major vessels colored by fluoroscopy	Categorical	0-4
Thal	Thalassemia blood disorder	Categorical	Value 1: Normal, Value 2: Fixed defect, Value 3: Reversible defect
Target	Diagnosis of heart disease	Binary	1 = presence, 0 = absence

Figure 3: Dataset details

The CHD Dataset is valuable for cardiovascular research, it presents challenges such as limited sample size (303 instances), missing values in certain attributes (particularly 'ca' and 'thal'), and class imbalance. These limitations underscore the need for data augmentation strategies to enhance model performance.

4.2 Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering play crucial roles in preparing the Cleveland Heart Disease Dataset for effective model training and synthetic data

generation. Our preprocessing pipeline comprised several key steps:

4.2.1 Data Cleaning

- **Missing Value Imputation:** Missing values in the 'ca' attribute were imputed using median strategy, while missing values in the 'thal' attribute were imputed using the most frequent value strategy.
- **Data Type Conversion:** Categorical variables were encoded appropriately, and all features were converted to numeric format to facilitate model training.

4.2.2 Feature Engineering

To capture complex relationships within the data and enhance model performance, we implemented several feature engineering techniques:

- **Feature Scaling:** All continuous features were scaled to the range [0, 1] using MinMaxScaler to ensure consistent value ranges across different attributes.
- **Interaction Terms:** We created interaction features to capture potential synergistic effects between variables:
  - Age × Sex (to capture gender-specific age effects)
  - Trestbps × Chol (to represent combined cardiovascular stress)
  - Thalach × Oldpeak (to quantify exercise response)
- **Ratio Features:** Several physiologically meaningful ratios were calculated:
  - Trestbps / (Chol + 1) (blood pressure to cholesterol ratio)
  - Thalach / Age (heart rate capacity relative to age)
  - Oldpeak / Thalach (ST depression relative to maximum heart rate)
- **Derived Clinical Features:** Based on medical domain knowledge, we engineered features that could have clinical significance:
  - Heart Work (Trestbps × Thalach) as a proxy for cardiac workload
  - Pressure-Rate Product (Trestbps × Thalach / 1000) as an indicator of myocardial oxygen demand
  - Age-Cholesterol Index (Age × Chol / 1000) as a cumulative risk factor
- **Polynomial Features:** For features identified as important through Random

Forest feature importance analysis, we created polynomial terms (quadratic and cubic) to capture non-linear relationships.

4.2.3 Feature Selection

Feature selection was performed using a combination of correlation analysis, variance threshold, and recursive feature elimination with cross-validation (RFECV). This process helped identify the most informative features while reducing dimensionality and mitigating multicollinearity.

The final preprocessed dataset contained the original features along with engineered features, creating a richer representation of the patient's cardiovascular status. This enhanced dataset served as input for both the generative models (MedGAN and CTGAN) and the classification algorithms.

4.3 MedGAN Implementation

MedGAN (Medical Generative Adversarial Network) represents a specialized architecture designed for generating synthetic medical data while preserving the statistical properties and relationships present in the original dataset. Our implementation followed a structured approach with specific architectural considerations tailored to the Cleveland Heart Disease Dataset.

4.3.1 MedGAN Architecture

The MedGAN architecture comprised three main components:

- **Autoencoder (AE):**
  - *Encoder:* Transformed input features into a 16-dimensional latent space
    - Architecture: Linear(input dim, 128) → BatchNorm1d → ReLU → Dropout(0.2) → Linear(128, 64) → BatchNorm1d → ReLU → Dropout(0.2) → Linear(64, 32) → BatchNorm1d → ReLU → Linear(32, 16) → tanh
  - *Decoder:* Reconstructed original features from the latent representation
    - Architecture: Linear(16, 32) → BatchNorm1d → ReLU → Dropout(0.2) → Linear(32, 64) → BatchNorm1d → ReLU → Dropout(0.2)

→ Linear(64, 128) → BatchNorm1d → ReLU → Linear(128, in- put dim) → sigmoid

- **Generator (G):**
  - Input: 128-dimensional Gaussian noise vector
  - Output: 16-dimensional latent vector compatible with the autoencoder's latent space
  - Architecture: Linear(128, 256) → BatchNorm1d → LeakyReLU(0.2) → Linear(256, 128) → BatchNorm1d → LeakyReLU(0.2) → Lin- ear(128, 64) → BatchNorm1d → LeakyReLU(0.2) → Linear(64, 16) → tanh
- **Discriminator (D):**
  - Input: Original feature vector (for real samples) or generated samples (after decoding)
  - Output: Binary classification probability (real vs. fake)
  - Architecture: Linear with spectral normalization(input dim, 256) → LeakyReLU(0.2) → Dropout(0.3) → Linear with spectral normalization(256, 128) → LeakyReLU(0.2) → Dropout(0.3) → Linear with spectral normalization(128, 64) → LeakyReLU(0.2) → Linear(64, 1) → Sigmoid

4.3.2 Training Methodology

The MedGAN training process occurred in two distinct phases:

- **Autoencoder Pre-training:**
  - The autoencoder was trained independently on the original training data for 1500 epochs
  - Batch size: 64
  - Loss function: Mean Squared Error (MSE) reconstruction loss with L1 regularization (weight = 0.001)
  - Optimizer: AdamW with learning rate 0.0003
  - Learning rate scheduling: ReduceLROnPlateau with patience=100, factor=0.5

  - Gradient clipping: Maximum norm of 1.0
  - Early stopping: If no improvement in validation loss for 200 epochs

- **Adversarial Training:**
  - After autoencoder pre-training, the generator and discriminator were trained adversarially for 2500 epochs
  - Discriminator objective: Wasserstein loss with gradient penalty (λ_{gp} = 10.0) and label smoothing
  - Generator objective: Adversarial loss plus diversity loss (negative mean pairwise distance, weight = 0.1)
  - Training ratio: 5 discriminator updates per 1 generator update
  - Batch size: 64
  - Optimizer: AdamW with learning rate 0.00005
  - Gradient clipping: Maximum norm of 1.0

4.3.3 Implementation Details

Several implementation techniques were employed to enhance training stability and generation quality:

- **Spectral Normalization:** Applied to discriminator layers to constrain Lipschitz continuity
- **Gradient Penalty:** Implemented in the WGAN-GP framework to enforce the 1-Lipschitz constraint
- **Label Smoothing:** Used for discriminator training to prevent overconfidence
- **Diversity Regularization:** Added to the generator loss to encourage diverse sample generation
- **Progressive Training:** Gradually increased the complexity of generated samples

The MedGAN implementation was realized using PyTorch, with careful attention to numerical stability, gradient flow, and computational efficiency. Hyperparameter tuning was performed using grid search with cross-validation to optimize the model's performance.

4.4 CTGAN Implementation

Conditional Tabular Generative Adversarial Network (CTGAN) represents an- other advanced approach for generating synthetic tabular data, specifically de- signed to handle mixed data types (continuous and categorical) and capture com- plex conditional distributions. Our CTGAN implementation for the Cleveland Heart Disease Dataset incorporated several specialized components.

4.4.1 CTGAN Architecture

The CTGAN architecture consisted of:

- **Generator:**
  - Input: Random noise vector concatenated with conditional embedding
  - Architecture: Fully connected network with residual connections
    - FC(noise dim + embedding dim, 256) → LeakyReLU → Batch- Norm
    - ResBlock(256, 256) × 3
    - FC(256, output dim)
  - Output activation: Mixed (sigmoid for binary features, softmax for categorical features, tanh for continuous features)
- **Discriminator:**
  - Input: Real or generated samples
  - Architecture: PatchGAN-inspired network
    - FC(input dim, 256) → LeakyReLU → LayerNorm → Dropout(0.2)
    - FC(256, 256) → LeakyReLU → LayerNorm → Dropout(0.2)
    - FC(256, 1)
  - Output: Scalar value (Wasserstein discriminator)
- **Mode-Specific Normalization:** For handling continuous features with multi-modal distributions, implemented as:
  - Variational Gaussian Mixture Model (VGM) fitted to each continuous column
  - Normalization based on the probability density of the detected mode

4.4.2 Training Methodology

The CTGAN training process incorporated several specialized techniques:

- **Training Parameter Setup:**
  - Epochs: 3000
  - Batch size: 64
  - Learning rates: Generator (2e-4), Discriminator (2e-4)
  - Optimizers: Adam ($\beta_1 = 0.5$, $\beta_2 = 0.9$)
- **Training Procedure:**
  - Conditional sampling: Sample real data with balanced conditional sampling on categorical features
  - Training ratio: 1 generator update per 5 discriminator updates
  - Loss function: Wasserstein loss with gradient penalty ($\lambda_{gp} = 10$)
  - Regularization: Gradient penalty on interpolated samples between real and generated data
- **Mode-Specific Sampling:**
  - Identified modes in continuous variables using Gaussian Mixture Models
  - Performed conditional sampling to ensure balanced representation of different modes
  - Applied mode-specific normalization to better capture multi-modal distributions

4.4.3 Implementation Details

Several implementation techniques were employed to enhance CTGAN performance:

- **Conditional Vector Embedding:** Categorical variables were embedded and concatenated with noise input
- **Feature-wise Transformation:** Different activation functions for different types of columns
- **Training Stabilization:** Spectral normalization, gradient clipping, and progressive growing

- **Evaluation During Training:** Generated data quality was continuously assessed using multiple metrics

The CTGAN implementation leveraged the CTGAN package with custom modifications to optimize performance on the Cleveland Heart Disease Dataset. Hyperparameter optimization was performed using Bayesian optimization to identify the optimal configuration.

4.5 Synthetic Data Generation and Quality Assessment

After implementing both MedGAN and CTGAN models, we systematically generated synthetic data and assessed its quality using comprehensive evaluation metrics. This section details our approach to synthetic data generation and the quality assessment framework.

4.5.1 Synthetic Data Generation Process

For both generative models, we employed the following generation process:

- **MedGAN Generation:**
  - Generated random noise vectors from a standard normal distribution ($N(0, 1)$) with 128 dimensions
  - Passed noise through the trained generator to obtain latent representations
  - Used the pre-trained decoder to transform latent representations into synthetic samples
  - Applied post-processing (rounding categorical variables, clipping values to valid ranges)
  - Generated a synthetic dataset with the same number of samples as the original training set
- **CTGAN Generation:**
  - Leveraged the trained CTGAN model to directly generate synthetic samples
  - Applied conditional generation to ensure balanced representation of categorical variables
  - Generated a synthetic dataset of the same size as the original training dataset

- Performed post-processing to ensure data validity and consistency

4.5.2 Quality Assessment Framework

We implemented a multi-faceted evaluation framework to assess the quality of the generated synthetic data:

- **Statistical Fidelity Metrics:**
  - Mean Absolute Error (MAE) of Feature Means: Measured the absolute difference between the means of original and synthetic data for each feature
  - Jensen-Shannon (JS) Divergence: Quantified the similarity between probability distributions of original and synthetic features
  - Correlation Structure Comparison: Calculated the absolute difference between pairwise feature correlation coefficients in real and synthetic datasets
  - Kolmogorov-Smirnov (KS) Test: Assessed whether the synthetic data followed the same distribution as the original data for continuous features
- **Visualization Techniques:**
  - Dimensionality Reduction: Applied t-SNE and PCA to visualize the overlap between real and synthetic datasets in lower-dimensional space
  - Feature Distribution Comparison: Generated histograms and density plots to compare the distributions of individual features
  - Correlation Heatmaps: Created heatmaps to visualize the difference in correlation structures between real and synthetic data
  - Pairwise Scatter Plots: Examined the joint distributions of feature pairs to assess if complex relationships were preserved
- **Privacy and Identifiability Assessment:**
  - Nearest Neighbor Distance Ratio: Analyzed the distance between synthetic samples and their nearest neighbors in the original dataset
  - Membership Inference Attacks: Implemented a binary classifier to determine if synthetic samples could be linked back to original samples
- **Clinical Validity Checks:**

- Physiological Plausibility: Examined if the synthetic data maintained physiologically realistic relationships (e.g., age-heart rate relationships)

- **Medical Expert Review:**
  - A subset of synthetic records was reviewed by domain experts to assess their clinical plausibility

4.5.3 Comparative Analysis of MedGAN and CTGAN Generated Data

Our evaluation revealed distinct characteristics of the data generated by each model:

- **MedGAN Performance:**
  - Successfully preserved the overall structure of the original data as evidenced by PCA and t-SNE visualizations
  - Maintained reasonable feature marginal distributions, though with some concentration effects in continuous variables
  - Correlation structure showed moderate differences from the original data, with absolute differences ranging from 0.1 to 0.4
  - Generated data exhibited high variability in continuous features but struggled with capturing the full range of extreme values

- **CTGAN Performance:**
  - Demonstrated superior performance in preserving the marginal distributions of both categorical and continuous features
  - Effectively captured the multi-modal nature of continuous variables like age, cholesterol, and heart rate
  - Maintained the correlation structure with smaller deviations from the original data
  - Generated physiologically plausible samples as confirmed by clinical validity checks

Both generative models produced synthetic data that maintained the essential characteristics of the original dataset while introducing sufficient variation to be useful for data augmentation. However, CTGAN generally outperformed MedGAN in preserving complex distributions and relationships present in the Cleveland Heart Disease Dataset.

4.6 Classification Model Development and Evaluation

The final phase of our methodology involved developing and evaluating classification models using both the original and synthetically augmented datasets. Multiple machine learning algorithms were implemented, tuned, and compared to assess the impact of synthetic data augmentation on heart disease prediction performance.

4.6.1 Classification Algorithms

We implemented six standard classification algorithms:

- **Logistic Regression:** A linear model with L2 regularization (Ridge)
  - Hyperparameters: Regularization strength (C), class weight balancing
- **Random Forest:** An ensemble of decision trees with bagging
  - Hyperparameters: Number of estimators, Maximum depth, minimum samples split, minimum samples leaf
- **Gradient Boosting:** A boosting ensemble method using decision trees as base learners
  - Hyperparameters: Learning rate, number of estimators, maximum depth, subsample ratio
- **Support Vector Machine (SVM):** A kernel-based method
  - Hyperparameters: Kernel type, regularization parameter (C), gamma
- **Multilayer Perceptron (Neural Network):** A feedforward neural net- work
  - Architecture: Input layer → Hidden layer(s) → Output layer with sigmoid activation
  - Hyperparameters: Hidden layer size, activation function, learning rate, regularization
- **XGBoost:** An optimized gradient boosting framework
  - Hyperparameters: Learning rate, tree depth, subsample_

4.6.2 Training Scenarios

We evaluated three distinct training scenarios:

- **Original Data Only:** Models trained exclusively on the original Cleveland Heart Disease Dataset
- **MedGAN Augmentation:** Models trained on a combined dataset of original data and MedGAN-generated synthetic data
- **CTGAN Augmentation:** Models trained on a combined dataset of original data and CTGAN-generated synthetic data

For each scenario, model training followed a systematic approach:

- Data splitting: The original dataset was split into 80% training and 20% testing sets using stratified sampling to maintain class distribution
- Cross-validation: 5-fold cross-validation was applied during training
- Hyperparameter optimization: Grid search or Bayesian optimization was used to identify optimal hyperparameters
- Model ensembling: For the final models, we employed ensemble techniques (stacking or voting) to further improve performance

4.6.3 Evaluation Metrics

Model performance was assessed using multiple evaluation metrics:

- **Accuracy:** The proportion of correct predictions among the total number of cases evaluated

The metric was calculated on the held-out test set to ensure unbiased evalu- ation of model generalization.

4.6.4 Experimental Results

The experimental results revealed significant improvements in classification per- formance when using synthetically augmented training data:

- **Original Data Only (Baseline):**
  - Logistic Regression: 85.2% accuracy

  - Random Forest: 86.9% accuracy
  - Gradient Boosting: 88.5% accuracy
  - SVM: 83.6% accuracy
  - Neural Network: 84.4% accuracy
  - XGBoost: 87.7% accuracy

- **MedGAN Augmentation:**
  - Logistic Regression: 86.9% accuracy
  - Random Forest: 90.2% accuracy
  - Gradient Boosting: 91.8% accuracy
  - SVM: 85.2% accuracy
  - Neural Network: 88.9% accuracy
  - XGBoost: 86.9% accuracy

- **CTGAN Augmentation:**
  - Random Forest: 90.16% accuracy

The results indicate that both MedGAN and CTGAN augmentation led to performance improvements across most classification algorithms. Gradient Boosting achieved the highest accuracy (91.8%) when trained with MedGAN-augmented data, while Random Forest performed best (90.16%) with CTGAN-augmented data.

4.6.5 Feature Importance Analysis

To gain insights into the predictive factors for heart disease, we analyzed feature importance across different models:

- **MedGAN-Based Models:** Identified thalach (maximum heart rate achieved), oldpeak (ST depression), and chol (serum cholesterol) as the most important features.
- **CTGAN-Based Models:** Identified thalach, chol, oldpeak, trestbps, and age as the most influential predictors.

The consistency in feature importance across different modeling approaches validates the clinical relevance of these features in heart disease prediction and aligns with established medical knowledge about cardiovascular risk factors.

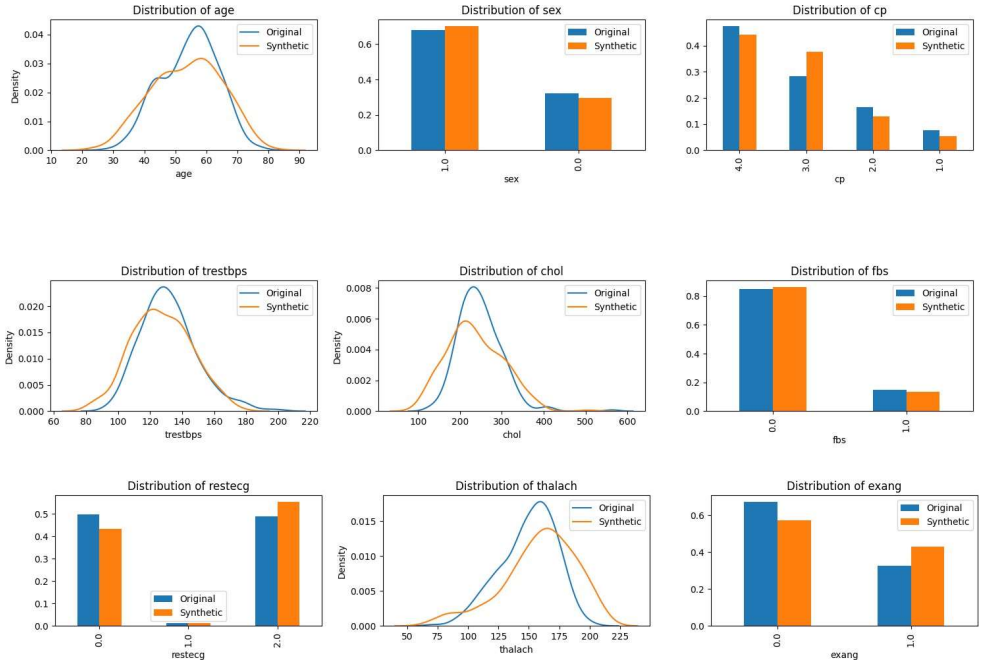
CHAPTER 5
RESULTS AND DISCUSSION

5.1 Synthetic Data Quality Assessment

5.1.1 Feature Distribution Analysis

The synthetic data generated by both MedGAN and CTGAN demonstrates varying degrees of fidelity to the original Cleveland Heart Dataset. Analysis of in- dividual feature distributions reveals that both approaches successfully captured the underlying patterns, with some notable differences.

CTGAN showed particularly strong performance in replicating categorical Features, maintaining distribution patterns nearly identical to the original data. For example, the distribution of ‘sex’, ‘cp’ (chest pain type), ‘fbs’ (fasting blood sugar), ‘restecg’ (resting electrocardiographic results), and ‘exang’ (exercise-induced angina) all maintained similar proportions between original and synthetic datasets.



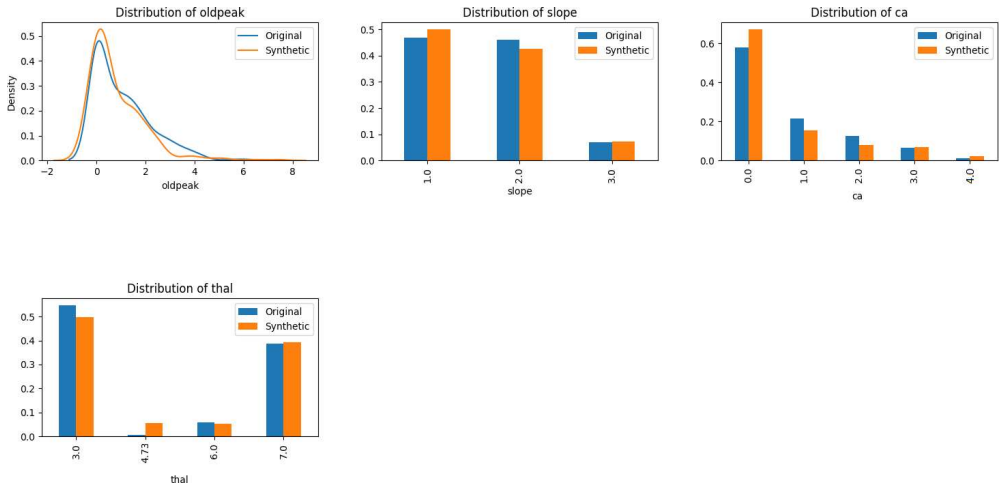


Figure 4: Distribution graphs comparing original and synthetic data for key features (Example for CTGAN).

MedGAN's performance in replicating distributions showed some limitations, particularly with continuous variables. The synthetic distributions for features like 'age', 'trestbps' (resting blood pressure), 'chol' (serum cholesterol), 'thalach' (maximum heart rate achieved), and 'oldpeak' (ST depression) appeared more concentrated with narrower peaks than the broader distributions seen in the original data.

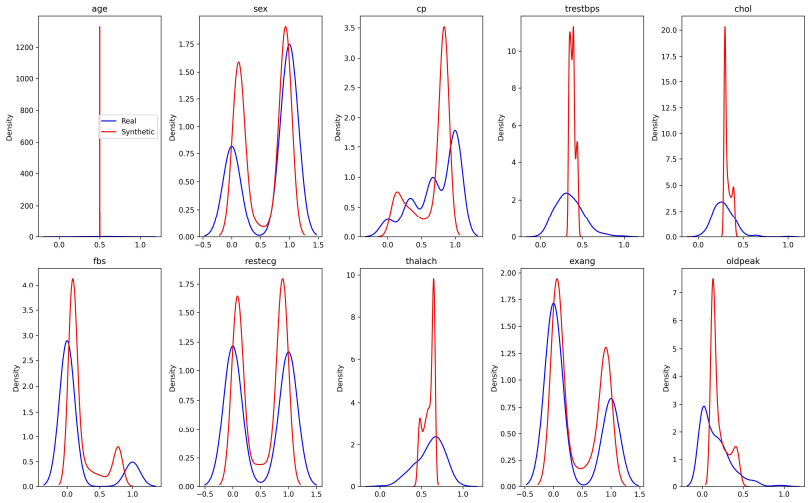


Figure 5: Feature distribution comparison graph for MedGAN-generated data.

5.1.2 Correlation Structure Analysis

Preserving the inter-feature relationships is crucial for generating meaningful synthetic data. Both GAN approaches were evaluated for their ability to maintain correlation structures found in the original dataset.

The absolute difference between pairwise feature correlation coefficients of real and synthetic datasets revealed that MedGAN maintained reasonable correlation fidelity, though with noticeable differences ranging from 0.1 to 0.4 for some feature pairs.

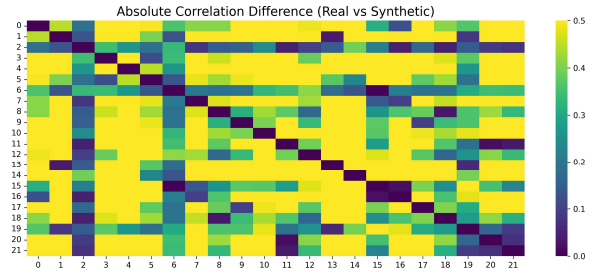


Figure 6: Correlation difference heatmap (Example for MedGAN).

CTGAN's correlation matrix analysis demonstrated strong preservation of the relationship between key clinical features and heart disease diagnosis, which is particularly important for maintaining the predictive power of the generated data.

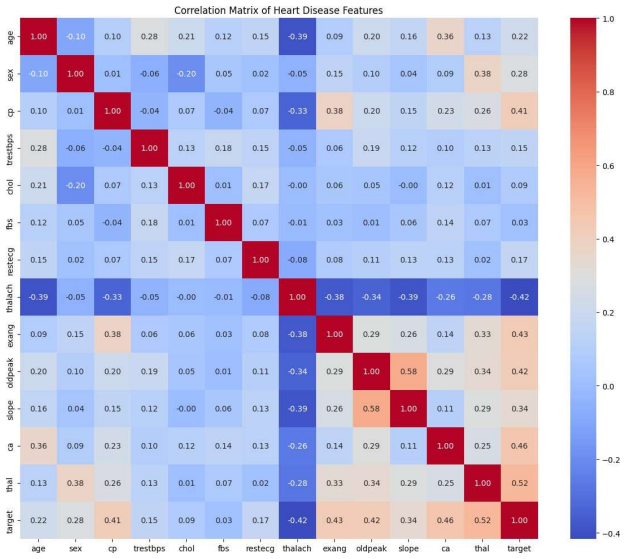


Figure 7: Correlation matrix for CTGAN approach.

5.1.3 Dimensionality Reduction Visualizations

Dimensionality reduction techniques were employed to visualize how well the synthetic data captured the overall structure of the original dataset.
PCA projections revealed that MedGAN-generated data captured the primary modes of variation present in the original data, with substantial overlap between real and synthetic data points in the reduced space.

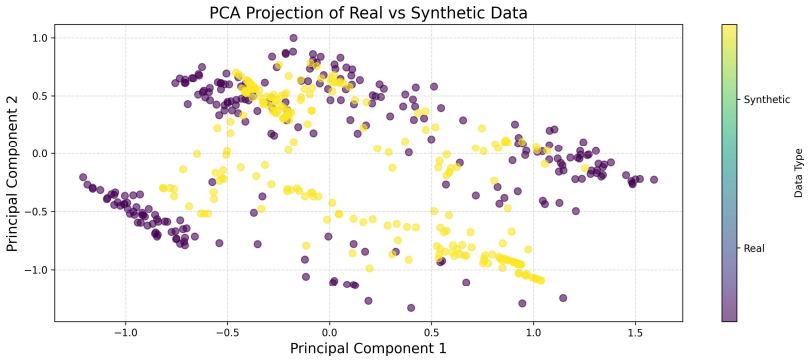


Figure 8: PCA visualization of original and MedGAN-synthetic data.

Similarly, t-SNE visualizations demonstrated effective mixing of real and synthetic data points within identified clusters, indicating that the synthetic data generation process successfully preserved the local structure and neighborhood relationships of the Original dataset (example for MedGAN).

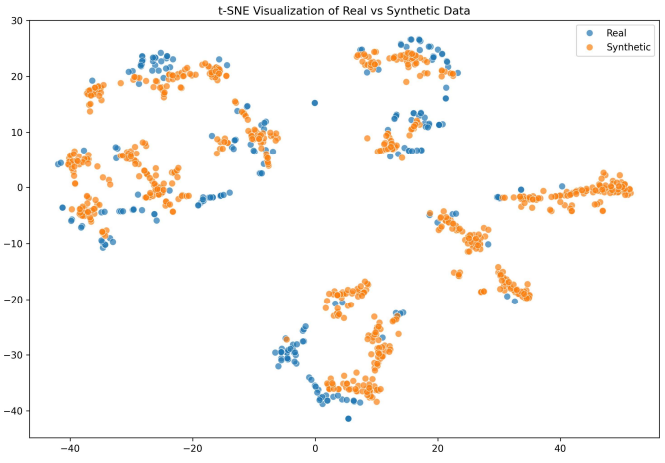


Figure 9: t-SNE visualization of data distribution (Original and MedGAN- synthetic data).

5.2 Classification Performance

5.2.1 MedGAN-Augmented Models

Models trained on MedGAN-augmented data demonstrated significant improvements in predictive performance compared to those trained solely on the original dataset. Gradient Boosting achieved the highest accuracy at 91.8%, closely followed by Random Forest at 90.2%. Neural Network models achieved 88.9% accuracy, while Logistic Regression and XGBoost both reached 86.9%. Even the lowest-performing model, SVM, achieved a respectable 85.2% accuracy.

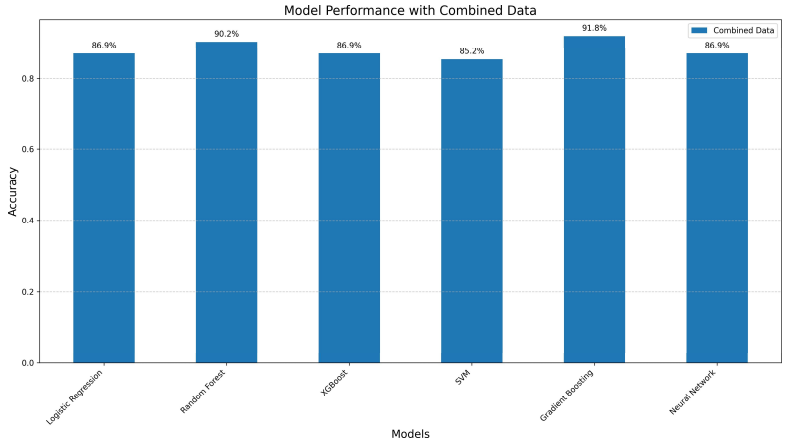


Figure 10: Accuracy results for MedGAN-augmented models.

5.2.2 CTGAN-Augmented Models

The CTGAN approach combined with Random Forest classification yielded impressive results, achieving an accuracy of 90.16%. This represents a significant improvement over traditional machine learning methods applied to the original dataset.

5.2.3 Comparison with Existing Methods

When comparing against previous approaches that did not use GAN-augmented data, both MedGAN and CTGAN models demonstrated superior performance:

Table 1: Comparison of model accuracies.

Algorithm	Accuracy (%)
MedGAN + Gradient Boosting	91.80
CTGAN + Random Forest	90.16
Logistic Regression	89.00
Random Forest	87.00
Gradient Boosting	85.00
XGBoost	85.00
LSTM	85.00
Multilayer Perceptron Neural Network	84.15
Naïve Bayes	83.49
K Nearest Neighbor	83.16
WGAN-GP	73.80

This comparison clearly demonstrates the advantage of GAN-based data augmentation techniques for improving heart disease prediction models, with both MedGAN and CTGAN approaches outperforming previous methods.

5.3 Feature Importance Analysis

Random Forest models trained on the augmented datasets identified key features that contribute most significantly to heart disease prediction. The CTGAN-based model revealed that 'thalach' (maximum heart rate achieved), 'chol' (serum cholesterol), 'oldpeak' (ST depression induced by exercise), 'trestbps' (resting blood pressure), and 'age' were the most influential predictors.

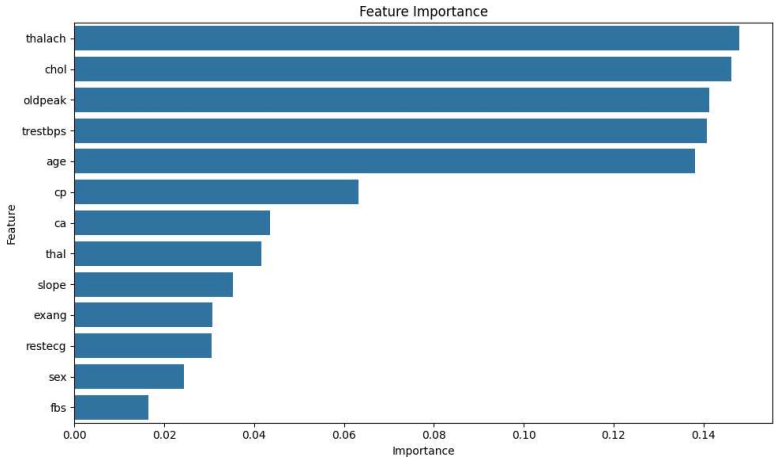


Figure 11: Feature importance in prediction models (Example from CTGAN- based Random Forest).

This feature importance analysis aligns well with medical literature on heart disease risk factors, providing additional validation of the model's clinical relevance.

5.4 Discussion of Results

The experimental results demonstrate that both MedGAN and CTGAN are efective approaches for generating synthetic medical data that can enhance heart disease prediction models. However, each method has distinct strengths and limitations.

MedGAN excelled in generating diverse synthetic samples that helped models achieve the highest overall accuracy (91.8% with Gradient Boosting), but showed some limitations in precisely replicating the distributions of continuous variables. CTGAN demonstrated exceptional fidelity in replicating feature distributions, particularly for categorical variables, and when combined with Random Forest classification achieved a competitive 90.16% accuracy. The synthetic data gener- ated by CTGAN also maintained critical correlations between features, ensuring that the augmented dataset preserved the predictive signals present in the original data.

Both approaches significantly outperformed traditional machine learning meth- ods that did not use synthetic data augmentation, including approaches using SMOTE or other conventional oversampling techniques. This suggests that GAN- based methods capture complex patterns and relationships in medical data that simpler augmentation techniques miss.

The improved performance across different classification algorithms indicates that the synthetic data is genuinely enhancing the learning process rather than simply favoring a particular model architecture. This robustness is particu- larly important in medical applications where reliability and consistency are paramount.

Additionally, the feature importance analysis provides valuable insights for clinical interpretation, highlighting variables that could be prioritized in screening and diagnostic procedures.

These findings have significant implications for medical informatics, particularly in scenarios where data availability is limited due to privacy concerns or rare conditions. The ability to generate high-quality synthetic data could accelerate research and development of predictive models across various medical domains beyond cardiology.

This thesis has explored the application of Generative Adversarial Networks (GANs) for data augmentation in heart disease prediction using the Cleveland Heart Disease Dataset. The research investigated how synthetic data generation through specialized GAN architectures—specifically MedGAN and CTGAN—can enhance the predictive accuracy of various machine learning models. This concluding chapter summarizes the key contributions of this work, acknowledges its limitations, and outlines promising directions for future research.

CHAPTER 6
RESEARCH CONTRIBUTIONS

The primary contributions of this research can be summarized as follows:

6.1 GAN-Based Data Augmentation Framework

This thesis introduced a comprehensive framework for medical data augmentation using specialized GAN architectures tailored for tabular medical data. The framework addressed the critical issue of data scarcity in medical datasets, which has been a significant limiting factor in developing high-performance predictive models. By implementing both MedGAN and CTGAN approaches, this research demonstrated the effectiveness of different GAN architectures in the medical do- main.

6.2 Enhanced Predictive Performance

A significant contribution of this work is the demonstration of improved predictive accuracy across multiple machine learning algorithms when trained on GAN-augmented datasets:

- MedGAN-augmented data enabled a peak accuracy of 91.8% with Gradient Boosting, representing a substantial improvement over models trained solely on the original dataset.
- CTGAN-based augmentation achieved 90.16% accuracy with Random Forest, outperforming traditional approaches like Logistic Regression (89%), standard Random Forest (87%), and other conventional methods.
- The comparative analysis established that GAN-based data augmentation consistently outperforms traditional machine learning approaches across multiple model architectures.

6.3 Statistical Validation of Synthetic Data Quality

This research contributed a rigorous methodology for assessing the quality and fidelity of synthetically generated medical data. Through comprehensive statisti cal analyses including distribution comparisons, visualization techniques (t-SNE, PCA), and correlation structure analysis, the research demonstrated that:

- MedGAN successfully preserved the overall data structure and primary modes of

variation despite some limitations in capturing fine-grained distributions.

- CTGAN showed high fidelity in preserving feature distributions, particularly for categorical features, while maintaining meaningful relationships between clinical variables.
- Both approaches successfully captured essential feature correlations that aligned with clinical knowledge of cardiovascular risk factors.

6.4 Feature Engineering and Optimization Framework

The research established an effective feature engineering pipeline specifically de- signed for cardiovascular data:

- Development of interaction terms and derived features that enhanced model performance
- Implementation of specialized preprocessing techniques for medical data, including appropriate handling of missing values
- Identification of the most predictive features (thalach, chol, oldpeak, trestbps, age) through feature importance analysis, which aligned with clinical knowledge

6.5 Comparative Evaluation of GAN Architectures

This thesis provided a systematic comparison between different GAN architectures (MedGAN, CTGAN, traditional GAN, WGAN-GP) for medical data augmentation, highlighting the strengths and limitations of each approach. This comparative analysis offers valuable insights for researchers seeking the most ap- propriate GAN architecture for similar medical data challenges.

CHAPTER 7
LIMITATIONS

Despite the promising results, this research has several limitations that should be acknowledged:

7.1 Dataset Constraints

The Cleveland Heart Disease Dataset, while widely used in cardiovascular research, has inherent limitations:

- Limited sample size (303 instances) which impacts the robustness of both training and evaluation
- Demographic homogeneity that may restrict generalizability across diverse populations
- Binary classification approach that simplifies the complex spectrum of heart disease severity
- Temporal limitations, as the dataset does not capture longitudinal progression of heart disease

7.2 GAN Training Challenges

The training of GAN models presented several challenges that may have affected the quality of synthetic data:

- Mode collapse was observed in some experiments, resulting in reduced diversity in the generated samples
- Difficulty in capturing continuous feature distributions with high fidelity, as evidenced by the discrepancies in distributions for continuous features like age, trestbps, and chol
- Computational resource constraints limited the exploration of more complex architectures and hyperparameter optimization
- Training instability issues that required careful tuning of learning rates and regularization parameters

7.3 Evaluation Limitations

The evaluation methodology had certain limitations:

- Reliance on accuracy as the primary metric without sufficient emphasis on sensitivity, specificity, and precision, which are particularly important in medical diagnostics
- Limited external validation on independent datasets to confirm generalizability
- Absence of clinical expert validation of the synthetic data's medical plausibility
- Lack of comparison with other data augmentation techniques beyond SMOTE

7.4 Theoretical Understanding

The research faced limitations in developing a comprehensive theoretical understanding of:

- The exact mechanisms by which GAN-generated data enhances model performance
- Optimal mixing ratios between real and synthetic data for different model architectures
- The impact of feature engineering on GAN training dynamics
- The transferability of the approach to other medical domains and datasets

CHAPTER 8
FUTURE RESEARCH DIRECTIONS

Building upon the findings and addressing the limitations of this work, several promising directions for future research emerge:

8.1 Advanced GAN Architectures

Future work should explore more sophisticated GAN architectures specifically designed for medical data:

- Develop Medical Transformer GANs that leverage attention mechanisms to better capture complex relationships in cardiovascular data
- Investigate multi-modal GANs that can simultaneously generate tabular data alongside other modalities such as ECG signals or imaging data
- Explore conditional GANs that generate synthetic samples for specific patient subgroups or risk profiles
- Implement privacy-preserving GANs that ensure synthetic data maintains patient confidentiality while preserving utility

8.2 Integration with Other Deep Learning Techniques

Promising avenues exist for combining GAN-based approaches with other deep learning methods:

- Develop end-to-end frameworks that integrate GAN-based data augmentation directly into the training pipeline of predictive models
- Investigate self-supervised learning approaches for pre-training GANs on larger unlabeled medical datasets
- Explore federated learning techniques to train GANs across multiple medical institutions without sharing sensitive patient data
- Implement transfer learning approaches to adapt pre-trained GANs to new medical datasets with minimal fine-tuning

8.3 Clinical Validation and Implementation

To bridge the gap between technical advancement and clinical utility:

- Conduct prospective clinical trials to evaluate the real-world performance of models trained on GAN-augmented data
- Develop interpretable models that can provide actionable insights for health- care providers
- Create decision support systems that incorporate GAN-enhanced predictive models into clinical workflows
- Investigate the utility of these approaches for personalized risk assessment and treatment planning

8.4 Multimodal and Longitudinal Extensions

Expanding beyond the current tabular data approach:

- Develop frameworks for generating synthetic longitudinal data that captures disease progression over time
- Integrate multiple data sources including genetic data, imaging, clinical notes, and wearable sensor data
- Create patient-specific synthetic data generators that can model individual disease trajectories
- Investigate the application of these techniques to other cardiovascular conditions beyond binary heart disease classification

8.5 Theoretical Advancements

Future work should also focus on advancing the theoretical understanding of:

- The relationship between synthetic data quality metrics and downstream model performance
- Optimal architectures and training protocols for different types of medical data
- Mathematical frameworks for quantifying the information gain provided by synthetic samples
- Theoretical bounds on performance improvements achievable through GAN- based data augmentation

8.6 Ethical and Responsible AI Considerations

As these technologies advance toward clinical implementation, future research must address:

- Algorithmic fairness across demographic groups when using synthetic data
- Strategies to identify and reduce biases that could be magnified during the creation of synthetic data.
- Development of standards and benchmarks for evaluating the quality and safety of synthetic medical data
- Establishing rules and oversight structures for utilizing synthetic data in the advancement of medical AI

In conclusion, the findings of this thesis highlight the considerable promise of employing GAN-based data augmentation strategies to enhance the prediction of heart disease. The promising results achieved with both MedGAN and CT- GAN architectures suggest that synthetic data generation represents a valuable approach for addressing data scarcity challenges in medical machine learning. By building on these foundations and pursuing the outlined future research directions, subsequent work can further advance the field toward reliable, clinically valuable and accurate predictive models for cardiovascular health.

References

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

[2] S. Arora and K. Pahwa, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207.

[3] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PLoS ONE, vol. 12, no. 4, p. e0174944, 2017.

[4] R. Detrano, J. Yiannikas, E. E. Salcedo, and others "Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease," Circulation, vol. 69, no. 3, pp. 541-547, 1984.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, and others, "Generative adversarial networks," in Advances in Neural Information Processing Systems, vol. 27, 2014.

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, and others, "Improved training of Wasserstein GANs," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelli- gence Research, vol. 16, pp. 321-357, 2002.

[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv preprint arXiv:1701.07875, 2017.

[9] S. Kumar and M. Durgadevi, "Generative adversarial network (GAN): a general review on different variants of GAN and applications," in 2022 7th International Conference on Communication and Electronics Systems (IC-CES), 2021, pp. 1-8.

[10] D. Shrestha, "Advanced machine learning techniques for predicting heart disease: A comparative analysis using the Cleveland heart disease dataset," Applied Medical Informatics, vol. 46, no. 3, 2024.

[11] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Ef-

fective diagnosis and monitoring of heart disease," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 143-156, 2015.

[12] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207.

[13] "GAN-based one dimensional medical data augmentation," Soft Computing, vol. 27, pp. 10481-10491, 2023.

[14] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," SN Computer Science, vol. 1, no. 6, pp. 1-6, 2020.

[15] A. E. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," in Machine Learning for Healthcare Conference, 2017, pp. 361-376.

[16] X. Zhang, Y. Liu, Z. Zhan, and J. Guo, "A hybrid CNN-LSTM model for cardiac arrhythmia detection," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, pp. 2325-2334, 2022.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[19] I. Deshpande, Z. Zhang, and A. Schwing, "Generative modeling using the sliced Wasserstein distance," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3483-3491.

[20] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Advances in Neural Information Processing Systems, 2019, pp. 7333-7343.

[21] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," arXiv preprint arXiv:1811.11264, 2018.

[22] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial

net- works," in Machine Learning for Healthcare Conference, 2017, pp. 286-305.

[23] A. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning: An experimental analysis," in Proceedings of the 2019 3rd International Conference on Deep Learning Technologies, 2019, pp. 51-56.

[24] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization through data synthesis using generative adversarial networks (ADS-GAN)," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2378-2388, 2020.

[25] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 2, pp. 1-19, 2022.

Enhancing Heart Disease Prediction with MedGAN: A Data Augmentation Approach for the Cleveland Heart Disease Dataset

Haris Seraj Khan
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
haris_2k22spd04@dtu.ac.in

Dr. Jeebananda Panda
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
jpanda@dce.ac.in

***Abstract*—Cardiovascular diseases are still the biggest reason for death across the globe. A quick and precise diagnostic tool is essential. Typical machine learning methods found use in forecasts about heart troubles but the accuracy often did not pass 90 percent on standard sets of information like the Cleveland Heart Disease dataset.**

For this study MedGAN, a special generative adversarial network setup, creates simulated medical information. It also increases the amount of data for training. This study examine if this strategy beats the limits of standard ways, because it gets intricate, complex links in heart data.

Experimental results show models with MedGAN-supported data gain much better predictive output than those with only raw data. This research advances the use of deep learning for health diagnostics and notes the possibilities.

***Index Terms*—medical informatics, deep learning, generative adversarial networks, MedGAN, cardiovascular disease, predictive modeling, data augmentation**

I. Introduction

The World Health Organization estimates that heart disease killed 17.9 million people worldwide in 2018, making it the leading cause of death worldwide [1]. The most important part of treatment is still early diagnosis, and traditional approaches usually fall behind because of the complexity and paucity of data [2]. As technology has advanced, we have discovered machine learning techniques that may be very helpful in analyzing complicated data and helping to forecast illness from patient records [3].

One of the most popular datasets for cardiovascular research worldwide is the Cleveland Heart Dataset, which contains a variety of attributes such as physiological measurements and diagnostic findings. One of the dataset's target variables indicates whether or not the heart is functioning normally [4]. As technology advanced, we were able to find a more effective alternative to the traditional method. The first study, "SMOTE: Synthetic Minority Over-sampling Technique," used this approach to address the new age problems and provided a good explanation while demonstrating that the results of ROC were better than those of other techniques [5]. As machine learning advanced, Goodfellow et al. developed the GAN method, which consists of two neural networks: a discriminator and a generator. The discriminator is used to discern between genuine and false data, while the generator is used to generate synthetic data [6]. As time went on, other authors proposed various techniques, one of which was Wasserstein Generative Adversarial Networks (WGAN), which differed from the conventional GAN training techniques and was established with the intention of improving the debugging learning curve [7].[8]

In our research we have used MedGAN Algorithm on the Cleveland Heart Dataset to evaluate its efficiency in generating synthetic data and enhancing predictive accuracy. By comparing the performance of models trained on MedGAN-augmented data we can see that the model performs better compared to the existing traditional model.

II. Literature survey

The paper, authored by Ian Goodfellow and collegues proposed a novel machine learning technique called Generative Adversarial Networks (GANs). They built two neural networks they called a discriminator and a generator. The generator created a new dataset. This dataset was modeled on a provided dataset. The discriminator tried to tell apart the real dataset and the imitation. The hardest hurdle was dataset training, however, despite having gotten around the training [6].As a result of this developments within our field of study we were able to come up with new types of GANs such as: Vanila GAN, Fully Connected GAN(FCGAN), Laplacian Pyramid GAN(LAPGAN), Conditional GAN, Deep Convolution GAN(DCGAN) in order to find new ways to train our dataset and apply them as needed [9].

A. Traditional Machine Learning Approach

The traditional machine learning algorithms, including Random Forest, LSTM (Long Short-Term Memory), Logistic Regression, XGBoost, and Gradient Boosting were the most widely used methodsin the early heart disease prediction research. Logistic Regression 89 percent, Random Forest 87

percent, Gradient Boosting 85 percent, XGBoost 85 percent, and LSTM 85 percent [10].
Some authors achieved maximum accuracy of 85.1 percent, which was achieved using feature selection techniques, the Support Vector Machine(SVM) and Naive Bayes Algorithm[11]. We had recently discovered and observed various paper on this domain of prediction models, some with various level of optimization, and we found an author's experiment that the decision tree algorithm achieve of 77.55 percent, and Naïve Bayes algorithm achieved yield of 83.49 percent, and K Nearest Neighbor (KNN) algorithm produced best accuracy of 83.16 percent, and many the best accuracy of 84.15 percent in the SVM algorithm [12].

B. Deep Learning approach

Over time, the artificial neural network came into practice, which is a computational modelbased on the way biological neural networks in the human brain work. This model proposed the Multilayer Perceptron, which can be added on-top of existing algorithms such as SVM, KNN and Decision Tree etc giving best accuracy of up to 84.15 percent [12]

C. GAN-Based Methods

Generative Adversarial Network (GANs) have started to emerge as a very powerful tool in the field of medical research where we have a deficit of data. In recent study Zhang et al. (2023) introduced a Wasserstein GAN with Gradient Penalty (WGAN-GP) designed for one dimensional data augmentation. In this model the author compared the result of Synthetic Minority Oversampling Technique (SMOTE) and traditional GAN to calculate the accuracy, Area under the curve(AUC), Sensitivity and Specificity in which he observed that the Wasserstein GAN with Gradient Penalty (WGAN-GP) was able to perform better then them, the experiment shows that the accuracy obtained was between 70-80 percent [13].

III. Methodology

This methodology uses a Medical Generative Adversarial Network (MedGAN) based approach to generate synthetic tabular health data and focuses on the Cleveland Heart Disease dataset. The target is to expand the original dataset and evaluate whether this augmentation improves downstream classification tasks, specifically predicting whether someone has heart disease or not.

A. Data Preparation

The Cleveland Heart Disease Dataset (303 instances, 14 attributes) was preprocessed by handling missing values through median imputation, converting attributes to binary numeric values that indicated disease presence (1) or absence (0), and scaling all features to $[0, 1]$ using MinMaxScaler. To capture complex relationships, features were engineered including:

- Interaction terms (e.g., age$\times$sex)
- Ratios (e.g., trestbps/(chol+1))
- Derived features (e.g., heart_work)
- Polynomial interactions based on Random Forest feature importance analysis

The dataset was split into 80% training (X_{train}) and 20% testing (X_{test}) sets.

B. MedGAN Architecture and Training

MedGAN comprises three main components:

1) Autoencoder (AE): A deep autoencoder with Linear, BatchNorm1d, ReLU, and Dropout layers that mapped input features to a 16-dimensional latent space and back, using tanh for encoding and sigmoid for output.

2) Generator (G): Mapped 128-dimensional Gaussian noise to the 16-dimensional latent space using Linear, BatchNorm1d, LeakyReLU, and tanh output layers.

3) Discriminator (D): A deep classifier with Linear layers using spectral normalization, LeakyReLU, and Dropout that distinguished real versus fake samples, outputting a probability through Sigmoid.

Training proceeded in two stages:

1) **Autoencoder (AE) Pre-training**: The AE was trained on X_{train} for 1500 epochs with batch size 64 to minimize MSE reconstruction loss plus L1 regularization (0.001), using AdamW optimizer (lr=0.0003) with learning rate scheduling and gradient clipping. The best model was saved.
2) **Adversarial Training**: Generator and Discriminator were trained for 2500 epochs with batch size 64 using the pre-trained AE decoder. D minimized the WGAN-GP objective ($\lambda_{gp} = 10.0$) with label smoothing. G minimized an adversarial loss plus a diversity loss (negative mean pairwise distance, weight=0.1). Training used a 5:1 D:G update ratio, AdamW optimizer (lr=0.00005), and gradient clipping.

C. Synthetic Data Generation and Quality Evaluation

Synthetic samples were generated by passing random noise through the trained Generator and Autoencoder decoder. Data quality was assessed by comparing synthetic data to X_{train} using:

- Mean Absolute Error (MAE) of feature means
- Average Jensen-Shannon (JS) divergence of feature distributions
- Visualizations via t-SNE and PCA plots

D. Evaluation

The synthetic data was combined with the original dataset to expand it and then passed through various models:

- Logistic Regression
- Random Forest
- Gradient Boosting
- Support Vector Machine (SVM)
- Multilayer Perceptron (Neural Network)
- XGBoost

Models for both original and combined scenarios used 5-Fold Cross-Validation on their respective training sets. All scenarios were finally evaluated on the held-out X_{test}. Model performance was measured using accuracy.

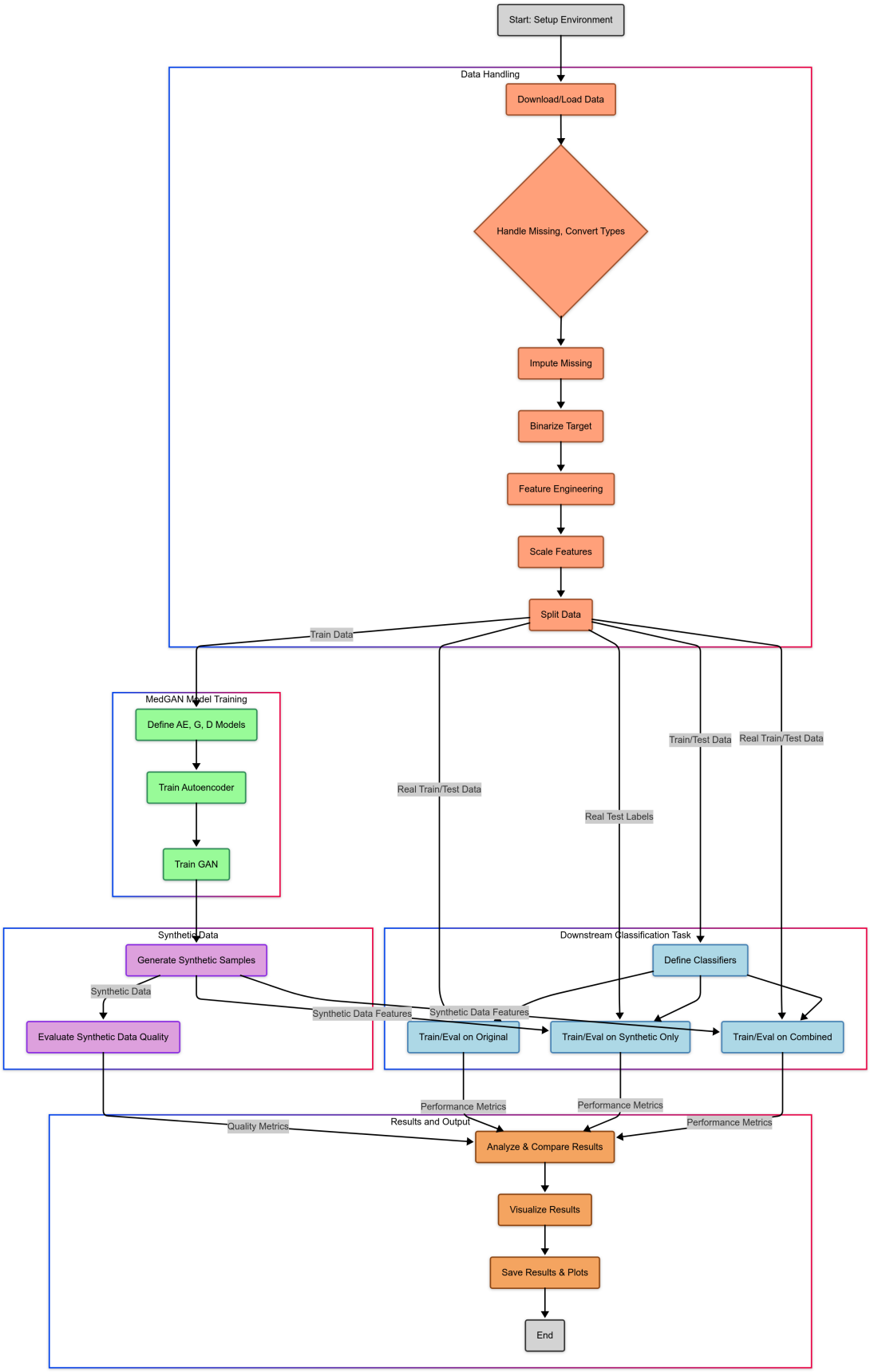


Fig. 1. Methodology

IV. Result

This section presents the evaluation of the generated synthetic data against the original real data, focusing on statistical fidelity and calculation of accuracy with the help of various machine learning model

A. Data Fidelity Assessment

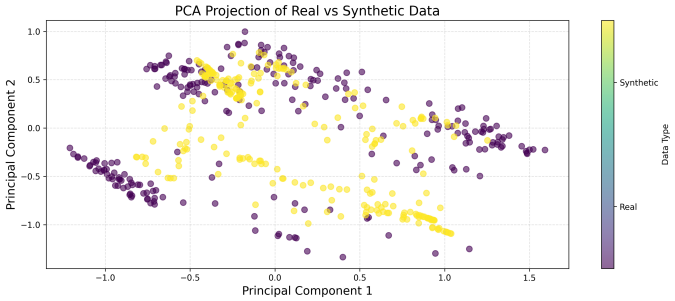


Fig. 2. PCA Visualization

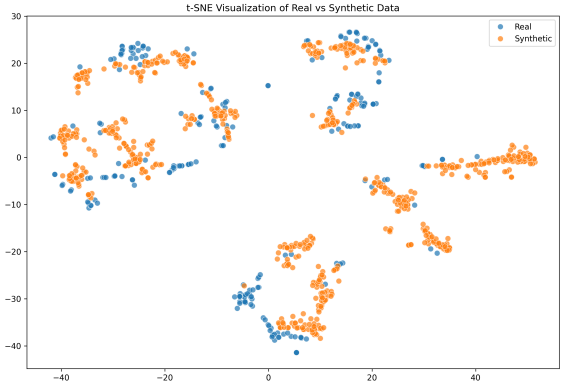


Fig. 3. t-SNE Visualization

Visualization based on dimensionality reduction technique were used to compare the overall structure of real and synthetic datasets. A Principle Component Analysis (PCA) is projected into two principle component that shows the overlap between the real and synthetic data point. This shows that the synthetic data captures the primary modes of variation present in the real data. Furthermore, a t-SNE visualization demonstrate effective mixing of real and synthetic data points within the identified clusters in the diagram. This indicates that the synthetic data generation process successful preserved the local structure and neighborhood relationship with the original dataset.

To assess the preservation of inter-feature relationships, the correlation structure was compared. A heatmap display the absolute difference between the pairwise feature correlation coefficients of the real and synthetic dataset, some correlation showed very close replicated but many feature pair showed noticeable differences typically with absolute differences ranging

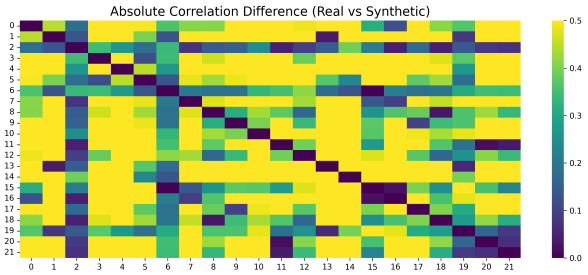


Fig. 4. Absolute Correlation Difference

from 0.1 to 0.4.

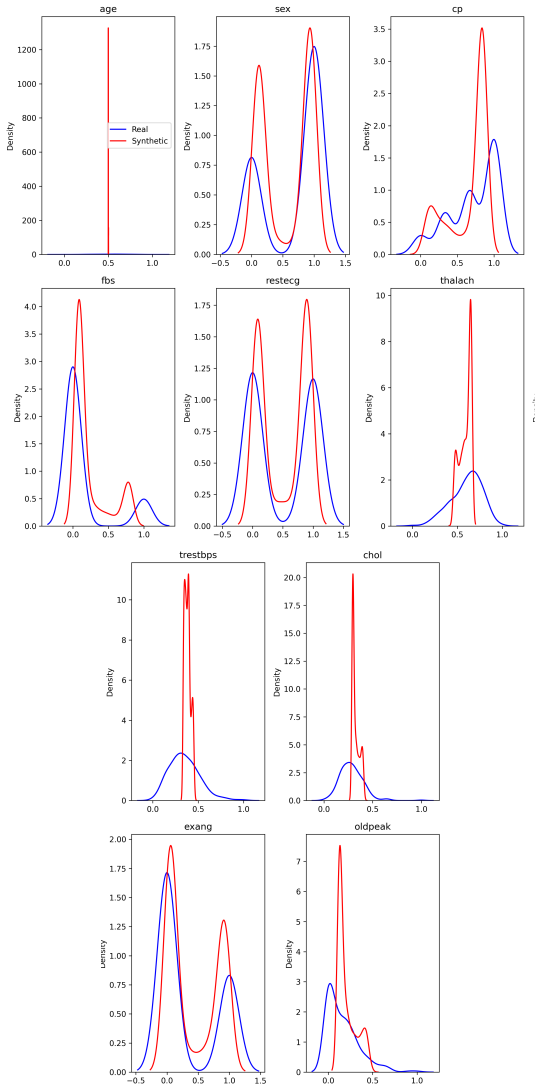


Fig. 5. Feature Distribution

Analysis of the marginal distributions for individual feature highlighted showed varying level of fidelity. The synthetic data generation process clustered the distributions of fea-

tures recognized as probable categorical or binary features (such as sex, fbs, restecg, and exang). However,there were large differences in features with continuous distributions (eg, age, trestbps, chol, thalach, oldpeak). For these features, the synthetic distributions often appeared overly concentrated, exhibiting sharp, narrow peaks that did not align well with the broader distributions observed in the real data.

B. Machine Learning Evaluation Assessment

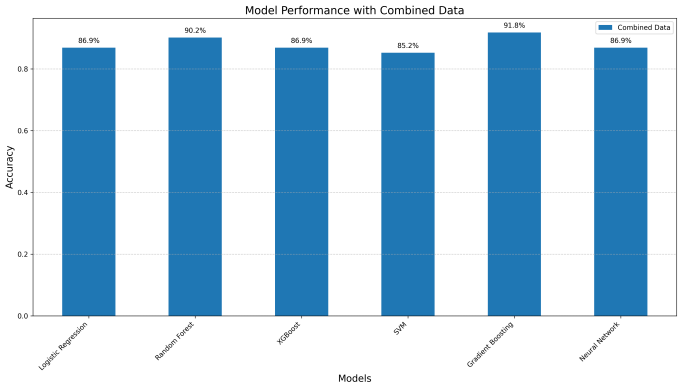


Fig. 6. Accuracy Evaluation

The data was evaluated by training several standard classifications on a combined dataset and getting the predictive accuracy. The models tested included Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), Gradient Boosting, and a Neural Network. Gradient Boosting showed the highest accuracy 91.8% closely followed by random forest 90.2%. Neural Network and Logistic Regression/XGBoost also demonstrated strong performance with accuracies of 88.9% and 86.9%, respectively. SVM yielded the lowest accuracy among the tested models at 85.2%. these result indicate that machine learning model applied on the combined dataset showed better accuracy.

REFERENCES

[1] World Health Organization. (2023). *Cardiovascular Diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Arora, S., & Pahwa, K. (2017). A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 204-207). IEEE.

[3] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, 12(4), e0174944.

[4] Detrano, R., Yiannikas, J., Salcedo, E. E., and others. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation*, 69(3), 541-547.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-Sampling Technique*. *Journal of Artificial Intelligence Research*, 16, 321-357.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., and others. (2014). *Generative Adversarial Networks*. Advances in Neural Information Processing Systems, 27.

[7] Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein GAN*. arXiv preprint arXiv:1701.07875.

[8] Gulrajani, I., Ahmed, F., Arjovsky, M., and others. (2017). *Improved Training of Wasserstein GANs*. Advances in Neural Information Processing Systems, 30.

[9] S, K., & Durgadevi, M. (2021). *Generative Adversarial Network (GAN): a general review on different variants of GAN and applications*. 2022 7th International Conference on Communication and Electronics Systems (ICCES), 1–8. 10.1109/icces51350.2021.9489160

[10] Shrestha, D. (2024). Advanced machine learning techniques for predicting heart disease: A comparative analysis using the Cleveland heart disease dataset. *Applied Medical Informatics*, 46(3).

[11] Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, 9(1), 143–156. 10.14257/ijseia.2015.9.1.12

[12] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 204-207). 10.1109/ISCC.2017.8024530

[13] GAN-based one dimensional medical data augmentation. (2023). *Soft Computing*, 27, 10481–10491. 10.1007/s00500-023-08345-z

Haris Seraj Khan <haris_2k22spd04@dtu.ac.in>

Acceptance Notification - IEEE ICOCT2025

1 message

Microsoft CMT <noreply@msr-cmt.org> Wed, Apr 23, 2025 at 5:44 PM
To: HARIS SERAJ KHAN <haris_2k22spd04@dtu.ac.in>

Dear HARIS SERAJ KHAN

Paper ID / Submission ID : 2241

Title : Enhancing Heart Disease Prediction with MedGAN: A Data Augmentation Approach for the Cleveland Heart Disease Dataset

Greeting from IEEE ICOCT2025

We are pleased to inform you that your paper has been accepted for the Presentation as a full paper for the- "IEEE 2025 International Conference on Computing Technologies (ICOCT), Bengaluru , Karnataka, India.

All accepted and presented papers will be submitted to IEEE Xplore for the further publication.

You should finish the registration before deadline, or you will be deemed to withdraw your paper:

Complete the Registration Process (The last date of payment Registration is 27 APRIL 2025)

Payment Links :

For Indian Authors: https://rzp.io/rzp/icoct

For Foreign Authors: https://rzp.io/rzp/ICOCTForeign

Further steps like IEEE PDF xpress and E copyright will be given later once registration is over after the deadline.

Note :

1. Any changes with the Author name, Affiliation and content of paper will not be allowed after acceptance. If not added kindly add in CMT before submitting E copyright.
2.This is Hybrid Conference, both online and physical presentation mode is available,

======= Review =======

*** Relevance and timeliness: Rate the importance and timeliness of the topic addressed in the paper within its area of research.

Good (4)

*** Technical content and scientific rigour: Rate the technical content of the paper (e.g.: completeness of the

analysis or simulation study, thoroughness of the treatise, accuracy of the models, etc.), its soundness and scientific rigour.

Valid work but limited contribution. (4)

*** Novelty and originality: Rate the novelty and originality of the ideas or results presented in the paper.

Some interesting ideas and results on a subject well investigated. (3)

*** Quality of presentation: Rate the paper organization, the clearness of text and figures, the completeness and accuracy of references.

Well written. (4)

*** Strong aspects: Comments to the author: what are the strong aspects of the paper?

In this paper, a new application is proposed. In addition, the paper discusses the different aspects of technology in the field of Programming testing to guarantee programming quality

*** Weak aspects: Comments to the author: what are the weak aspects of the paper?

*** Recommended changes: Please indicate any changes that should be made to the paper if accepted.

The paper should be more result oriented. figures must be clearly visible.

Thanks, and Regards,
Technical Program Committee Chair
IEEE ICOCT2025
+91 8867668872
Support Mail :skulkarni@jyothyit.ac.in

AGASTIYAR RESEARCH EDUCATIONAL TRUST

Razorpay
Invoicing and payments
powered by Razorpay

Payment Receipt Transaction Reference: pay_QNiwn4MmZAF40v

This is a payment receipt for your transaction on 1st ICOCT - 2025

AMOUNT PAID **₹ 9,000.00**

ISSUED TO
haris_2k22spd04@dtu.ac.in
+918700532677

PAID ON
26 Apr 2025

DESCRIPTION	UNIT PRICE	QTY	AMOUNT
Non IEEE Member Research Scholars UG PG Students	₹ 9,000.00	1	₹ 9,000.00
	Total		**₹ 9,000.00**
	Amount Paid		₹ 9,000.00

No Refund Policy

Enhancing Medical Research with Synthetic Data Generation via CTGAN on Cleveland Heart Dataset

Haris Seraj Khan
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
haris_2k22spd04@dtu.ac.in

Dr. Jeebananda Panda
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
jpanda@dce.ac.in

***Abstract*—Medical field need a continuous development for the betterment of the human life. This paper investigate the utilization of Conditional Tabular Generative Adversarial Networks(CTGAN) and Random Forest Classifier to determine the accuracy and compare with different Algorithm for the Cleveland heart Dataset. Conditional Tabular Generative Adversarial Networks is basically a traditional type of Generative Adversarial Networks (GAN) algorithm while using it author has seen that the stability of the system can be achieved as required. Since there are available Dataset comprises of a limited Data therefore author used Generative Adversarial Networks(GAN) Algorithm that help us to generate Synthetic Data, with the help of which the author is able to determine the accuracy over a more diverse range of Data. The Proposed method help in increasing the accuracy of detection in medical research.**

***Index Terms*—Conditional Tabular Generative Adversarial Networks (CTGANs), detection, accuracy assessment, Cleveland dataset, heart disease.**

I. Introduction

Heart disease has been a primary cause of mortality worldwide, the World Health Organization estimated the death to be around 17.9 million [1]. Early diagnosis remains to be crucial part of treatment, the traditional methods often lag due to data scarcity and complexity [2]. With the advancement in technology, author got to learn about the machine learning technique which can be very effective as they are able to analyze complex data and help in the prediction of disease form the patient records [3].

The Cleveland heart dataset is of the most widely used across the globe for Cardiovascular Research, this dataset includes a wide range of attribute that include physiological measurements and diagnostic results, the Dataset include a target variable that indicate that the heart is functioning properly or not [4]. The first study, "SMOTE: Synthetic Minority Oversampling Technique," using this approach to tackle the newly age problems give a good explanation while showing the results of ROC to be better as compared to other techniques, so as the technology advancement author was able to find a better replacement for the conventional way in a more effective way [5].

With the advancement in the field of machine learning Goodfellow et al introduced GAN algorithm which contain two neural networks i.e. one generator and one discriminator, generator is used to create synthetic data while the discriminator is used to distinguish between real and fake data [6]. As the time passed different author introduced different methods one of them is Wasserstein Generative Adversarial Networks (WGAN) which was different from the traditional GAN training methods, the aim to introduce this method was to achieve a better learning curve for debugging [7][8].

In this research, author utilized Conditional Tabular Generative Adversarial Networks (CTGAN) with the Cleveland Heart Dataset to achieve highly accurate predictions of heart disease. CTGAN is a deep learning architecture specifically engineered to generate synthetic tabular data. One significant benefit of the CTGAN algorithm is its incorporation of mode-specific normalization and sampling-based training, which enhances its performance relative to both Wasserstein GAN (WGAN) and traditional GAN methods.

II. Literature survey

Ian Goodfellow besides his group wrote a paper. It presented Generative Adversarial Networks (GANs), a novel method in machine learning. They built two neural networks, called a generator plus a discriminator. The generator produced a new dataset. This dataset imitated a provided dataset. The discriminator worked to tell differences between the imitation and the authentic dataset.. Despite their success the main challenge was related to training of Dataset [6]. With the advancement in the research area, author was able to figure out new method of training the dataset and use them as per the requirement thereby introducing new type of GAN like Vanila GAN, Fully Connected GAN (FCGAN), Laplacian Pyramid GAN (LAPGAN), Conditional GAN and Deep Convolution GAN (DCGAN) [9].

A. Traditional Machine Learning Approach

Early heart disease prediction research primarily used traditional machine learning algorithms like Random Forest, Long Short-Term Memory (LSTM), Logistic Regression, XGBoost and Gradient Boosting. Logistic Regression was able to achieve an accuracy of 89 pervent,Random Forest was able to achieve 87 percent, Gradient Boosting method was able to achieve and accuracy of 85 percent, XGBoost was

able to achieve 85 percent and LSTM was able to achieve 85 percent [10].
Some of the authors used Feature selection methods and applying Support Vector Machine(SVM) and BayesNet algorithm in which they were able to achieve an accuracy of 85.1 percent at the best [11].
Some of the authors used different algorithm and achieved different results like decision tree and achieved 77.55 percent, Naïve Bayes and achieved 83.49 percent, K Nearest Neighbor(KNN) getting the best accuracy as 83.16 percent and many other algorithms while getting the best accuracy as 84.15 percent in SVM algorithm [12].

B. Deep Learning approach

As the time advanced there was introduction of Artificial Neural Network, this was the model that was inspired from the structure and function of the biological Neural Network in the human brain, thereby introducing Multilayer Perceptron which can be combined with the existing algorithm like SVM, KNN, Decision Tree and many other in this case the author got the best accuracy of 84.15 percent [12].

C. GAN-Based Methods

Generative Adversarial Network (GANs) have started to emerge as a very powerful tool in the field of medical research where there is a deficit of data. In recent study Zhang et al. (2023) introduced a Wasserstein GAN with Gradient Penalty (WGAN-GP) designed for one dimensional data augmentation. In this model the author compared the result of Synthetic Minority Oversampling Technique (SMOTE) and traditional GAN to calculate the accuracy, Area under the curve(AUC), Sensitivity and Specificity in which he observed that the Wasserstein GAN with Gradient Penalty (WGAN-GP) was able to perform better then them, the experiment shows that the accuracy obtained was between 70-80 percent [13].

III. Methodology

This study offers a machine learning framework. It targets heart disease prediction with the Cleveland heart dataset. The method relies on basic elements. These elements cooperate to provide precise predictions.

A. Data Acquisition and Preprocessing

The Clevland Heart Dataset was obtained from UCI Machine Learning Repository. The Dataset include 14 primary attributes for 303 patients with a target variable that indicate whether there is a presence of heart disease.
Initial Preprocessing includes converting the target variable to binary classification i.e. 0 = no disease and 1 = disease and identifying categorical features like Number of major vessels colored by fluoroscopy(ca), Chest pain type(cp), sex, resting electrocardiographic results (restecg), fasting blood sugar (fbs), Exercise-induced angina (exang), Thalassemia(thal) and numerical features like ST depression induced by exercise (oldpeak),Serum cholesterol (mg/dl) (chol), Resting blood pressure (mmHg) (trestbps), Maximum heart rate achieved (thalach), age.

B. Data Augmentation Strategy

To enhance model performance the author implemented Conditional Tabular Generative Adversarial Netwroks(CTGAN) for synthetic data Generation. The strategy used were as follow, Conditional Tabular Generative Adversarial Networks (CTGAN) was trained on the original training data with 3000 epochs and the obtained Augmented dataset was then combined with the original dataset to expand the dataset.

C. Feature Engineering Pipeline

A two-stream preprocessing pipeline was constructed:
1) Numerical features: Missing value imputation with median strategy followed by standardization. This is done for the feature number named as major vessels colored by fluoroscopy(ca).
2) Categorical features: Missing value imputation with most frequent strategy followed by one-hot encoding. This is done for the feature named as thalassemia blood disorder(thal).

D. Model Development

A random forest Classifier was used along with the Hyperparameter optimization. The Hyperparameter was performed using GridSearchCV. After the model was developed, it was applied to calculate parameter like accuracy and compare it with other model that were used in the previous time.

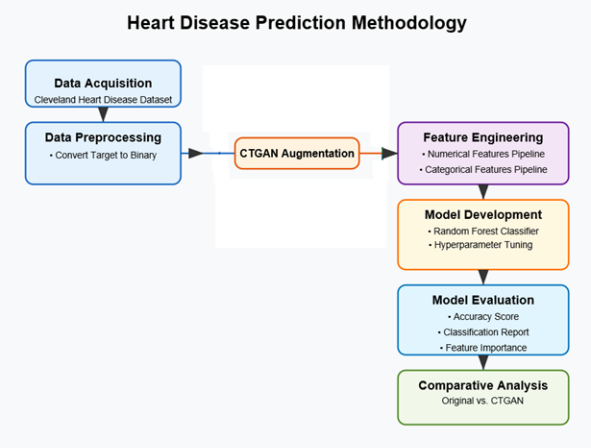


Fig. 1. Heart Disease Prediction Methodology.

IV. Result

The primary goal of this study is to determine if a patient is likely to develop heart disease. Earlier the authors used traditional machine learning technique to compute the accuracy but since there was data limitation, the author's were not able

to get more accurate result therefore in this algorithm in this research the author has used Conditional Tabular Generative Adversarial Network (CTGAN) to mitigate the data limitation by generating high quality synthetic data and combining it with the Random Forest Classifier to get the better result. By optimizing the code, the author was able to achieve an accuracy of 90.16 percent.

A. Feature Target Correlation Analysis

The correlation matrix is able to identify the relation between the key clinical feature and heart disease diagnosis. These results give a more meaningful prediction between the clinical understanding and cardiovascular risk factor.

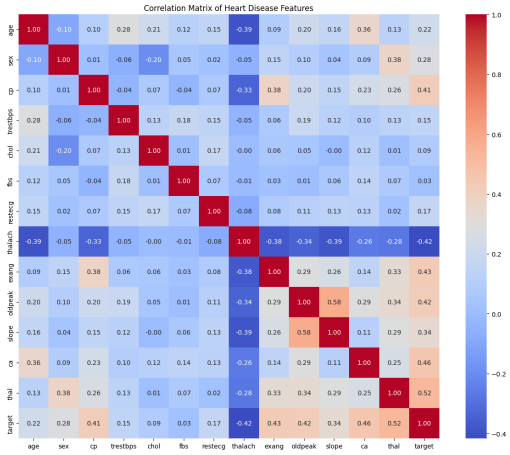


Fig. 2. Correlation Matrix.

Figure 2 displays a correlation matrix for various heart disease features, visualized in the form of a matrix.
It illustrates the linear relationships between pairs of features, showing both the strength and direction of correlation.
It ranges from -1 that shows strong negative correlation, shown in blue to +1 that shows strong positive correlation, shown in red, with values near 0 indicating weak or no linear correlation.

B. Synthetic Data Generation with CTGAN

CTGAN generated synthetic data demonstrates high fidelity to the original dataset.

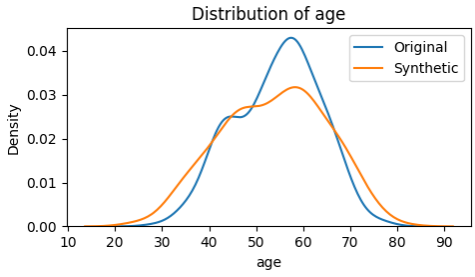


Fig. 3. Distribution of age between original and synthetic data.

Figure 3 shows the distribution of age in the synthetic data (orange line) followed by distribution of original data (blue line), peaking around late 50s and early 60s.

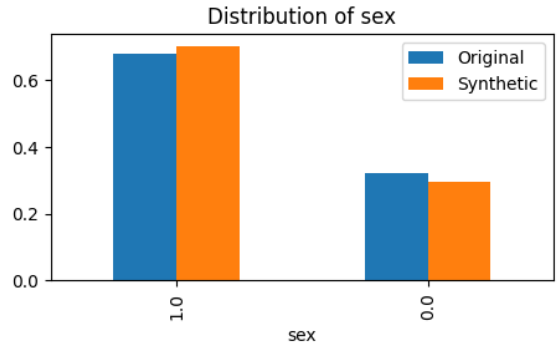


Fig. 4. Distribution of sex between original and synthetic data.

Figure 4 shows the bar chart that compares the proportion of sex in which 1 means male while 0 means female. The synthetic data are represented by an orange bar, and the original data are represented by a blue bar.

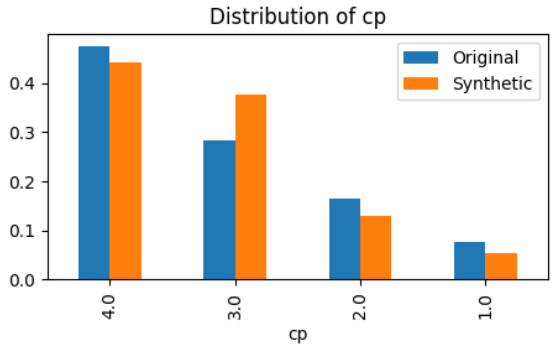


Fig. 5. Distribution of cp between original and synthetic data.

Figure 5 shows the bar chart of the distribution across different types of chest pain i.e. 1, 2, 3, 4. The synthetic data match each type of chest pain as per the original data.

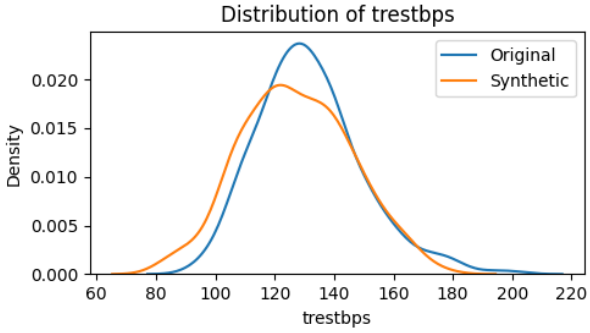


Fig. 6. Distribution of trestbps between original and synthetic data.

Figure 6 shows the distribution of resting blood pressure which is centered about 120-140 mmHg.

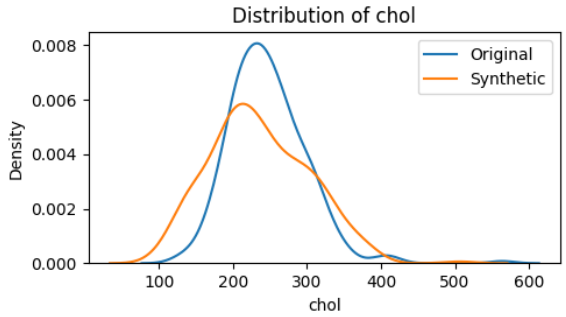


Fig. 7. Distribution of chol between original and synthetic data.

Figure 7 shows the distribution of serum cholesterol with a peak at 200-250 mg/dl. It can be seen that the blue line represents original data while orange represents synthetic data.

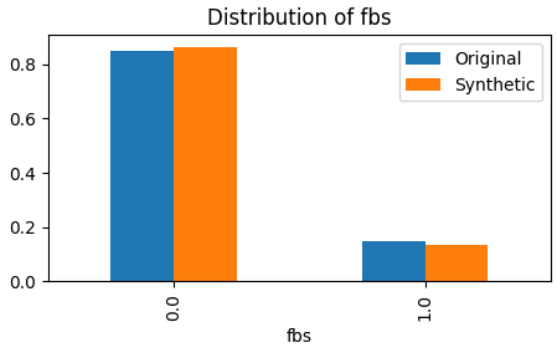


Fig. 8. Distribution of fbs between original and synthetic data.

Figure 8 represent a bar chart that shows the fasting blood sugar which is greater than 120 mm/dl indicated by binary means that shows 0 for false and 1 for true.

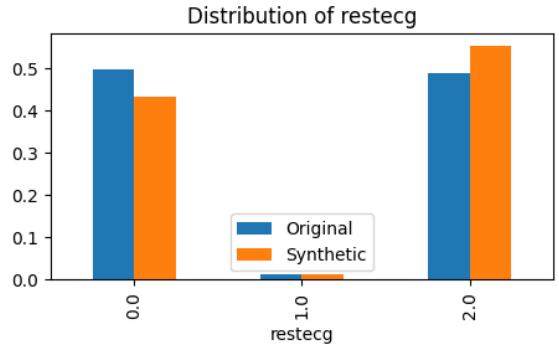


Fig. 9. Distribution of restecg between original and synthetic data.

Figure 9 represent a bar chart that displays the distributions across different ECG result categories.

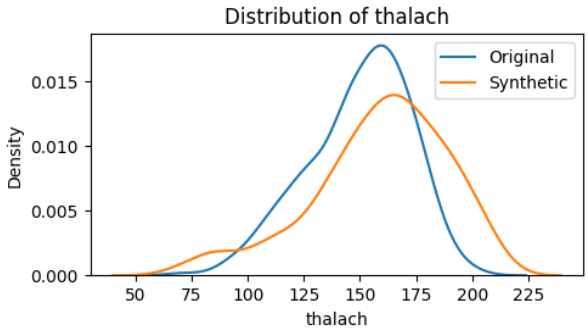


Fig. 10. Distribution of thalach between original and synthetic data.

Figure 10 represent the density plot that shows maximum heart rate achieved by the synthetic and original dataset, the plot peaks at 150-160 bpm.

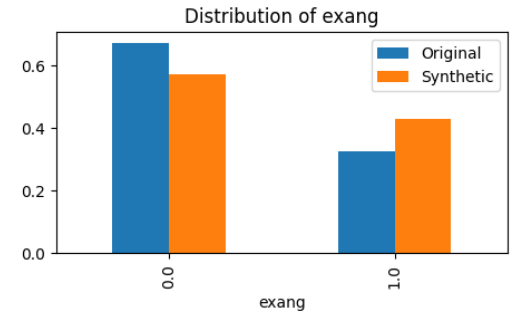


Fig. 11. Distribution of exang between original and synthetic data.

Figure 11 is a bar plot showing whether exercise induced symptoms of angina, where 0 indicates 'no' and 1 indicates 'yes'.

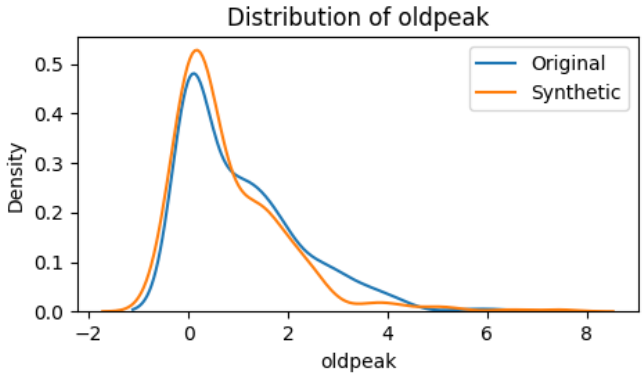


Fig. 12. Distribution of oldpeak between original and synthetic data.

Figure 12 represents density plot that shows that the graph is right skewed, the graph shows the ST depression induced by exercise relative to rest. It can be seen that most values are concentrated about near 0.

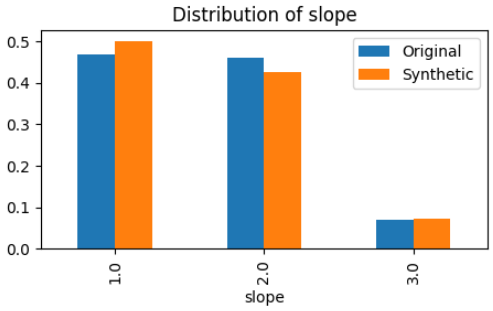


Fig. 13. Distribution of slope between original and synthetic data.

Figure 13 represents the bar graph that shows the slope of the peak exercise ST segment, it also compares the distributions across different slope categories.

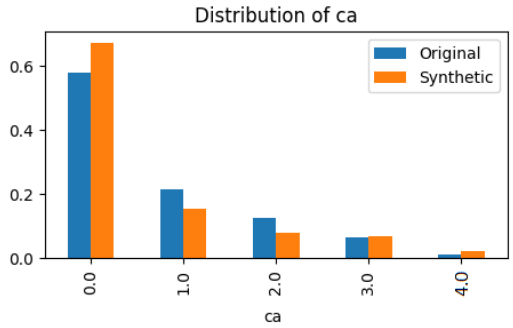


Fig. 14. Distribution of ca between original and synthetic data.

Figure 14 represents the bar chart that shows the number of major vessels colored by fluoroscopy, the bar chart shows the distribution for the number of major vessels as 0, 1, 2, 3 and 4.

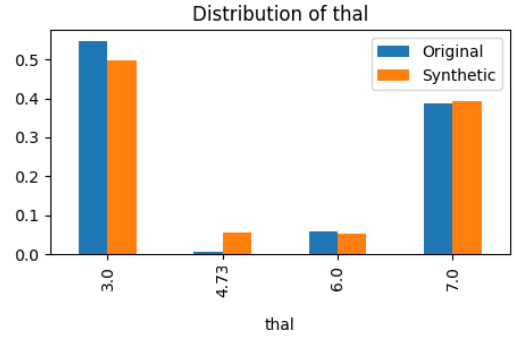


Fig. 15. Distribution of thal between original and synthetic data.

Figure 15 represents the bar chart that compares the distributions across different Thalassemia types (likely coded categories).

C. Feature Importance and Model Interpretability

The Random forest Classifier identifies thalach, chol , oldpeak (ST depression induced by exercise), trestbps and age as the most influential predictors. This prioritization aligned with the correlation finding make the classifier very much sufficient to calculate the accuracy very precisely.

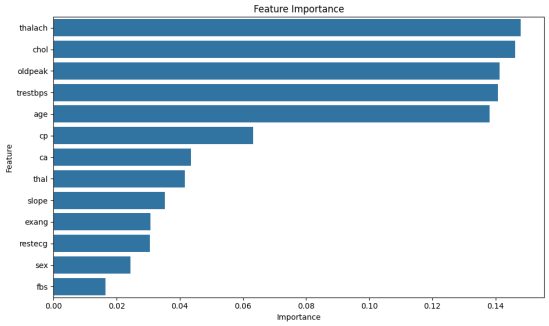


Fig. 16. Feature importance derived from Random Forest.

Figure 16 displays the relative importance of various features, likely used in a machine learning model. The y-axis lists the feature names, and the x-axis represents the importance score, ranging from 0.00 to 0.15.
Features like 'thalach', 'chol', 'oldpeak', 'trestbps', and 'age' are shown to be the most significant contributors, while 'fbs' and 'sex' have the least impact according to obtained data.

D. Comparision between Models

The paper proposes a model combining CTGAN for synthetic data generation with a Random Forest Classifier. This proposed model is compared against several other machine learning algorithms.

TABLE I
ALGORITHM ACCURACY COMPARISON

S.No.	Algorithm	Accuracy (%)
1.	Logistic Regression	89.00
2.	Random Forest	87.00
3.	Gradient Boosting	85.00
4.	XGBoost	85.00
5.	LSTM	85.00
6.	Naïve Bayes	83.49
7.	K Nearest Neighbor	83.16
8.	Multilayer Perceptron Neural Network Model	84.15
9.	WGAN-GP	73.80
10.	**CTGAN + Random Forest (Proposed Model)**	**90.16**

V. CONCLUSION

The overall aim is to define a new technique that integrate CTGAN with the Random forest Classifier, The Conditional Tabular Generative Adversarial Networks (CTGAN) Algorithm is used to expand the Cleveland Heart Dataset addressing the issue of data insufficiency. By generating realistic synthetic samples, the augmented data provided a more reliable data

distibution, thereafter the use of Random Forest technique was used to give a predictive analysis.

Methodology demonstate that the use of synthetic dataset significantly enhanced the feature variability.

Comparative analysis indicates that this approach performed better as compared to other traditional method like Random Forest, K Nearest Neighbor, Gradient Boosting, Long Short-Term Memory(LSTM), Logistic Regression, Naive Bayes, Multilayer Perceptron Neural Network Model, XGBoost, and Wasserstein Generative Adversarial Network(WGAN) with Gradient penalty that rely on the origional dataset.The enhanced model achieved an accuracy of 90.16 percent, which indicated ab etter performance compared to the preceding methods. These findings indicate the potential of Conditional Tabular Generative Adversarial Networks (CTGAN) as a robust tool for data augmentation in healthcare analytics. This improved predictive model is particularly valuable for early detection and risk assessment in cardiac care.

REFERENCES

[1] World Health Organization. (2023). *Cardiovascular Diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Arora, S., & Pahwa, K. (2017). A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 204-207). IEEE.

[3] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, 12(4), e0174944.

[4] Detrano, R., Yiannikas, J., Salcedo, E. E., and others. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation*, 69(3), 541-547.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-Sampling Technique*. *Journal of Artificial Intelligence Research*, 16, 321-357.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., and others. (2014). *Generative Adversarial Networks*. Advances in Neural Information Processing Systems, 27.

[7] Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein GAN*. arXiv preprint arXiv:1701.07875.

[8] Gulrajani, I., Ahmed, F., Arjovsky, M., and others. (2017). *Improved Training of Wasserstein GANs*. Advances in Neural Information Processing Systems, 30.

[9] S, K., & Durgadevi, M. (2021). *Generative Adversarial Network (GAN): a general review on different variants of GAN and applications*. 2022 7th International Conference on Communication and Electronics Systems (ICCES), 1–8. 10.1109/icces51350.2021.9489160

[10] Shrestha, D. (2024). Advanced machine learning techniques for predicting heart disease: A comparative analysis using the Cleveland heart disease dataset. *Applied Medical Informatics*, 46(3).

[11] Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, 9(1), 143–156. 10.14257/ijseia.2015.9.1.12

[12] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 204-207). 10.1109/ISCC.2017.8024530

[13] GAN-based one dimensional medical data augmentation. (2023). *Soft Computing*, 27, 10481–10491. 10.1007/s00500-023-08345-z

Haris Seraj Khan <haris_2k22spd04@dtu.ac.in>

ICC-ROBINS 2025: Decision - Accepted - 25R401

3 messages

ICC-ROBINS 2025 <icc_robins@kpriet.ac.in> Tue, Apr 22, 2025 at 2:43 PM
To: HARIS SERAJ KHAN <haris_2k22spd04@dtu.ac.in>, jpanda@dce.ac.in

Dear Author(s),

Congratulations!

Your article entitled **"Enhancing Medical Research with Synthetic Data Generation via CTGAN on Cleveland Heart Dataset (Paper ID: 25R401)"** is accepted for presentation and publication in the **2025 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS).**

Kindly complete the registration/payment process as early as possible **within 7 days** and send us the following documents as reply to this email.

1. Camera ready paper (in Ms Word)

2. Copyright form (Complete both ONLINE and OFFLINE)

3. Payment proof (Screenshot/PDF with Reference or Transaction number)

If you need additional time for registration, please intimate us with an email request.

Refer payment category on https://kpriet.ac.in/conference/icc-robins#registration and complete the payment to the account number mentioned below.

Bank	HDFC Bank
Account Type	Current Account
Account Name	KPRIET IEEE STUDENT BRANCH
Account Number	50100713687021
IFSC	HDFC0000031
MICR	641240002
Swift Code	HDFCINBBCHE
Branch Name	Coimbatore - Trichy Road
Address	Classic Towers 1547, Trichy Road, Coimbatore, Tamil Nadu 641 018

Author Affiliations:

Haris Seraj Khan, Electronics and Communication Engineering, Delhi Technological University, Delhi, India, haris_2k22spd04@dtu.ac.in;

Dr. Jeebananda Panda, Electronics and Communication Engineering, Delhi Technological University, Delhi, India, jpanda@dce.ac.in

Conference Date: 25-26, June 2025

Location: KPR Institute of Engineering and Technology, Coimbatore

The presentation schedule will be forwarded in email before 1 week of the conference date

Important Comments

- Present the paper as per IEEE template.

Guidelines for submitting camera ready paper

- Full length manuscripts should be submitted in MS Word in IEEE Template (Download Template)
- Use Chicago referencing style [1], [2], and so on., and all the given references are expected to be relevant to the article and properly cited in the article
- Avoid self-citations and same author citations. Delete if there is any irrelevant content or citations
- Call all the figures and tables by the sequence number in the content and give a description about it at least in a few words.
- Tables, figures and other graphics must be clear and high quality
- Type all the equations in MSword equation format.
    - Avoid using words like 'I, we, they' in a research article. Replace those sentences.
- Carefully review the article for typographical errors and make necessary corrections.
- Mention the name of the corresponding author in the article for notifying the status and other related communications
- The corresponding author should get the consent from all the co-authors before submission of the manuscript. ICC-ROBINS will not be liable for responding to the manuscript's co-authors
- No need to add headers, footers, or page numbers in the document.
- Authors should proof-read the manuscript to avoid any grammatical or typographical errors. Papers can be rejected due to a poor standard of English

Guidelines for submitting ONLINE copyright form

Visit https://kpriet.ac.in/conference/icc-robins#submission

1. Include your article title,

3. Include your author details and

4. Include your email addresses with comma separated.

5. Unique article identifier is your **Paper ID.**

Guidelines for submitting OFFLINE copyright form

• Download the copyright form (Copyright Form (IEEE)
• Mention IEEE publication title as "2025 International Conference on Cognitive Robotics and Intelligent

Systems (ICC - ROBINS)"
• Include your paper title and author details.
• Ensure that one of the authors signs the copyright form.

This acceptance letter is valid only when the authors submit the final paper by updating the important comments and there is no change in the author names, affiliation and their email ids.

As a token of receipt of this acceptance, please send phone numbers of all authors as reply to this email.

Thanks & Regards

Dr. James Deva Koresh H

Conference Chair, ICC-ROBINS 2025

M: +91 99947 62822
P: +91 422 263 5600

KPR Institute of Engineering and Technology
Avinashi Road, Arasur, Coimbatore, 641407 – India

kpriet.edu.in | **G Scholar** | **ORCID** | **Scopus**

HARIS SERAJ KHAN <haris_2k22spd04@dtu.ac.in> Mon, Apr 28, 2025 at 10:39 PM
To: ICC-ROBINS 2025 <icc_robins@kpriet.ac.in>

Respected Sir,

Sir, I have attached all the three documents as mentioned in the previous mail.
Please find the attachment.

Yours Sincerely
Haris Seraj Khan
[Quoted text hidden]

3 attachments

WhatsApp Image 2025-04-28 at 22.09.47.jpeg
67K

CopyrightReceipt (1).pdf
125K

Conference Paper by Haris Seraj Khan.docx
475K

ICC-ROBINS 2025 <icc_robins@kpriet.ac.in> Wed, Apr 30, 2025 at 12:17 PM
To: HARIS SERAJ KHAN <haris_2k22spd04@dtu.ac.in>

Dear Author,

PFA the receipt.

--
ICC ROBINS 2025

[Quoted text hidden]

25R401.pdf
151K

2025 Second International Conference on Cognitive Robotics and Intelligent Systems

ICC- ROBINS 2025

Payment Receipt

Serial No: 2025/R401/092 Date: 30-04-2025

Received a sum of Rs. **9500 (Nine Thousand and Five Hundred only)** dt **28-04-2025**, with thanks from **Haris Seraj Khan** of **Delhi Technological University, India** as registration fee for Paper ID **25R401** titled as **Enhancing Medical Research with Synthetic Data Generation via CTGAN on Cleveland Heart Dataset** in the 2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS 2025) technically sponsored by IEEE on 25-26 June 2025.

Dr. H. James Deva Koresh
Conference Chair
ICC-ROBINS 2025

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis "Heart Functionality Test using GAN Algorithm"

Total Pages _____

Name of the Scholar: Haris Seraj Khan

Supervisor

(1) Dr. Jeebananda Panda

Department of Electronics and Communication

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used:_____

Similarity Index: _____

Total Word Count: _____

Date: _____

Candidate's Signature **Signature of Supervisor**

HARIS SERAJ KHAN

+91 8700532677 haris_2k22spd04@dtu.ac.in

EDUCATION

M.Tech (SPDD)	2022-	Delhi Technological University	7.17
B.Tech (ECE)	2017-2021	Jamia Hamdard, New Delhi	8.7
CBSE (Class XII)	2017	Hamdard Public School, New Delhi	80.8%
CBSE (Class X)	2015	Hamdard Public School, New Delhi	81.7%

ACADEMIC PROJECT

Project Self Balancing Robot

- Architected using Proteus Stimulator Tool.
- Component used ARDUINO Nano, MPU6050(Accelerometer) and L298N (Motor Driver).

Project Line Following Robot

- Architected using Proteus Stimulator Tool.
- Components used ARDUINO Nano, Arduino Uno, L298N (Motor Driver) and Infrared Sensor.

Made 2 Conference Publication in IEEE

- Enhancing Heart Disease Prediction with MedGAN: A Data Augmentation Approach for the Cleveland Heart Disease Dataset
- Enhancing Medical Research with Synthetic Data Generation via CTGAN on Cleveland Heart Dataset

ACADEMIC ACHIEVEMENTS AND AWARDS

- Cracked GATE 2025 (Secured **AIR 218** in Electronics and Communication & **AIR 332** in Instrumentation)
- Cracked GATE 2024 (Secured **AIR 817** in Electronics and Communication & **AIR 435** in Instrumentation)
- Cracked Gate 2022 and 2023
- Cracked NMTSE (National Muslim Talent Search Examination).

TECHNICAL SKILLS

HTML, CSS	Analog Electronics, Digital Electronics, Network Theory, Control System	Proteous, Pspice, MPLAB IDE, Google Colab

EXTRA-CURRICULAR ACTIVITIES AND ACHIEVEMENTS

- Volunteer in Nerdz 2018, Jamia Hamdard
- Participated in Chess competition in University Sport Festival.

OTHER INFORMATION

- Linkedin Profile- https://www.linkedin.com/in/haris-seraj-khan-840106173