# Advancing Heart Failure Prediction: A Comparative Study of Traditional Machine Learning, Neural Networks, and Stacking Generative AI Models


Howard H. Nguyen *
Data Science Department
Harrisburg University
Pennsylvania, USA
info@howardnguyen.com

Maria Vaida, Ph.D. *
Data Science Department
Harrisburg University
Pennsylvania, USA
mvaida@harrisburgu.edu

Kevin Purcell, Ph.D.
Data Science Department
Harrisburg University
Pennsylvania, USA
kpurcell@harrisburgu.edu

Kevin Huggins, Ph.D.
Data Science Department
Harrisburg University
Pennsylvania, USA
khuggins@harrisburgu.edu

* Corresponding author

Srikar Bellur, Ph.D.
Data Science Department
Harrisburg University
Pennsylvania, USA
sbellur@harrisburgu.edu

Roozbeh Sadeghian, Ph.D.
Data Science Department
Harrisburg University
Pennsylvania, USA
rsadeghian@harrisburgu.edu



***Abstract*—Heart failure (HF) poses critical global health challenges, emphasizing the need for robust predictive models to support early diagnosis and enhance patient outcomes. Traditional machine learning (ML) models, such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting Machines (xGBM), have shown effectiveness but face limitations in handling nonlinear relationships, addressing class imbalances, and generalizing across datasets. Deep learning (DL) models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at identifying complex patterns but are hindered by computational requirements and limited interpretability, restricting clinical adoption. This research evaluates predictive models using nine datasets ranging from 299 to 400,000 records. Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalances, while a Stacking Generative AI (Gen AI) model was developed. This hybrid model integrates Generative AI with RF, GBM, and CNNs, enhancing underrepresented subgroup representation through synthetic data generation. The Stacking Generative AI model demonstrated superior performance, achieving 98% accuracy and a Receiver Operating Characteristic Area Under the Curve (ROC AUC) of 0.999 on a 1,025-record dataset. These results highlight the model's ability to handle complex data, enhance predictive accuracy, and improve clinical relevance. A web application further illustrates its practical value, offering an accessible platform for HF risk assessment. This study underscores the innovative role of hybrid models in advancing healthcare decision-making and improving patient care.**



***Keywords—machine learning, deep learning, neural networks, stacking models, generative AI.***


## I. Introduction

Heart failure (HF) is a significant public health issue due to its high morbidity and mortality rates, requiring early detection for improved patient outcomes and reduced healthcare burdens. Predictive models play a critical role in enabling timely and informed decision-making (Davis & Smith, 2023). Machine learning (ML) and deep learning (DL) techniques have shown substantial promise in healthcare, particularly for predictive tasks (Breiman, 2001; LeCun et al., 2015).

Traditional ML models, such as Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (xGBM), have demonstrated success in predictive applications. However, their inability to capture nonlinear and temporal complexities in healthcare data limits their performance. In contrast, neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) excel at identifying intricate patterns but face challenges in computational efficiency and interpretability, making them less ideal for clinical use (LeCun et al., 2015; Cho et al., 2014).

Hybrid stacking models offer a solution by combining the strengths of multiple algorithms to improve accuracy and generalizability. These models use a meta-learner to integrate predictions from base models, enhancing performance (Sagi & Rokach, 2018). This study introduces a Stacking Generative AI (Gen AI) model, which integrates GANs, RF, GBM, xGBM, and CNNs for HF prediction (Goodfellow et al., 2014). GANs address the challenge of class imbalances by generating synthetic data, which improves model performance on datasets with underrepresented minority cases (Frid-Adar et al., 2018; Yi et al., 2019).

This research evaluates traditional ML models, DL models, standalone Generative AI, and the proposed Stacking Generative AI model across nine datasets. Key questions include: (1) How do ML models compare to DL models like CNN and RNN? (2) What are the most influential predictors of HF? (3) Can a hybrid stacking model combining ML, DL, and GANs outperform single models? (4) How does incorporating Generative AI improve model performance? (5) What contributions can the Stacking Generative AI model make to HF prediction?

Initial findings indicate that the hybrid approach consistently outperforms standalone models. On smaller datasets of 1,000 and 1,025 records, the model achieved 98% accuracy and a ROC AUC of 0.999, effectively addressing class imbalance and capturing complex data patterns. By integrating advanced AI techniques, this research demonstrates the

potential of hybrid models to enhance HF prediction and support personalized care.

Finally, this study explores the limitations of synthetic data, such as biases that affect model generalizability. By evaluating the Stacking Generative AI model on nine datasets, including the Framingham dataset of 4,240 records, it investigates biases, performance variability, and real-world applicability.

## II. Literature Review

### A. Traditional ML Approaches in HF Prediction

Traditional ML models like RF, GBM, xGBM, and LR have been widely used in HF prediction due to their robust performance, but they often struggle with nonlinear relationships, class imbalances, and high-dimensional data. Chicco and Jurman (2020) identified RF as a top performer in HF survival prediction with 74% accuracy and an ROC AUC of 0.80, though its limited dataset restricted generalizability. Singh et al. (2024) achieved a 95.3% accuracy and a 0.97 ROC AUC by training a DNN on 5,888 records with advanced preprocessing techniques, but data complexity remained a challenge.

Optimization methods like Bayesian tuning and genetic algorithms have enhanced ML models, as shown by Rimal et al. (2024), who reported 89% accuracy with RF. Ensemble approaches further improved performance; Hasan and Saleh (2021) applied stacking to the Framingham dataset (4,239 records), achieving a 96.69% accuracy and 0.98 ROC AUC. However, these models did not integrate DL or Generative AI techniques, limiting scalability. The proposed Stacking Generative AI model addresses these limitations by incorporating synthetic data to tackle class imbalance and improve performance, achieving a 95% accuracy and a 99% ROC AUC.

### B. Neural Network-Based Approaches

DL models have advanced HF prediction by capturing complex data patterns missed by traditional ML models. Mahmud et al. (2023) introduced a lightweight metamodel combining ML algorithms, achieving 87% accuracy on 920 records. While efficient, it lacked the sophistication of advanced DL models. RNNs, particularly with GRUs, have shown promise in temporal modeling; Choi et al. (2017) achieved a ROC AUC of 0.883 using RNNs on EHR data. However, the absence of hybrid strategies limited broader applicability.

CNNs have also proven effective. Arooj et al. (2022) achieved 91.7% accuracy on a 1,050-record dataset using a DCNN, though the lack of generalizability across datasets remained a limitation. Emerging approaches like transformers have demonstrated potential in HF prediction. Sakthi et al. (2024) achieved 88.6% accuracy using transformers to identify heart anomalies, while Tuli et al. (2020) proposed HealthFog, an IoT-based framework integrating ensemble DL with fog computing, achieving a 91.2% accuracy and 0.94 ROC AUC. Despite scalability, reliance on device resources hindered broader clinical adoption.

### C. Hybrid and Stacking Models in HF Prediction

Hybrid models leverage the strengths of multiple algorithms to improve accuracy and generalizability. Ali et al. (2020) combined wearable sensor data with EMRs in a DL-based system, achieving a 98.5% accuracy, though the exclusive reliance on DL limited robustness. Mienye et al. (2020) achieved 93% accuracy with ensemble ML models but excluded DL methods, restricting the ability to capture complex data patterns.

Wankhede et al. (2022) integrated DL with the Tunicate Swarm Algorithm, achieving 97.5% accuracy on the Cleveland dataset, though its small size and lack of ML integration hindered scalability. Liu et al. (2022) utilized stacking with multiple classifiers, reporting ROC AUCs of 0.95 and 0.92 across two datasets. However, the absence of Generative AI techniques limited the ability to address class imbalance.

### D. Generative AI and GAN Frameworks in HF Prediction

GANs have emerged as effective tools for addressing class imbalance and data complexity in HF prediction. Khan et al. (2024) combined ML and DL with GANs to generate synthetic data, achieving 96.1% accuracy and a 0.927 ROC AUC. Anbarasu and Suruli (2022) introduced a deep ensemble learning model combined with GAN-based semi-supervised training, achieving accuracies of 86.54%, 84.83%, and 86.72% on the SPECT, WDBC, and Hallmarks datasets, respectively. Their approach used GANs to generate synthetic data and integrate multiple classifiers with a deep neural network.

Yu et al. (2024) proposed a GAN framework with a feature-enhanced loss function, achieving 94.62% accuracy and a 0.958 ROC AUC on the KORA cohort dataset. Similarly, Bhagawati and Paul (2024) achieved 93% accuracy and a 0.953 ROC AUC for coronary artery disease prediction using GANs.

The proposed Stacking Generative AI model builds on these advancements by synthesizing balanced datasets to improve minority class predictions. Its superior performance, achieving 95% accuracy and a 99% ROC AUC across nine datasets, demonstrates the potential of hybrid models in addressing data imbalance and advancing HF prediction.

## III. Methodology

### A. Stacking Generative AI Models

The Stacking Generative AI model integrates traditional ML models (RF, GBM, xGBM) with DL architectures (CNNs, RNNs) and GAN-generated synthetic data to address class imbalance and enhance predictive accuracy. This proposed hybrid framework adapts to dataset sizes, leveraging ML for smaller datasets and DL for larger, complex datasets. As a result, the model achieved a 98% accuracy and a 99.9% ROC AUC on a 1,025-record dataset and 96% accuracy with a 0.99 ROC AUC on a 400,000-record dataset.

### B. Overview of Methodology

The methodology involved preprocessing nine HF datasets (299 to 400,000 records) with techniques like data cleaning, normalization (Z-score), and SMOTE for class balancing. GANs further improved data robustness and generalizability by generating high-quality synthetic samples. Hyperparameter optimization using Grid Search Cross-Validation refined model

performance. Evaluation metrics included accuracy, ROC AUC, precision, recall, and F1-scores, consistently showing superior results for the Stacking Generative AI model.

### C. Data Collection and Preprocessing

Nine diverse datasets ensured the robustness and scalability of the model:

1) *299-Record Dataset (Pakistan):* Collected at the Faisalabad Institute of Cardiology, this dataset includes patients aged 40-95 years, focusing on clinical measures like ejection fraction and serum creatinine.
2) *303-Record Dataset (Cleveland, USA):* Derived from the UCI repository, it captures key attributes like chest pain type and serum cholesterol.
3) *1,000-Record Dataset (India):* Features clinical parameters like blood pressure and fasting blood sugar.
4) *1,025-Record Dataset (Global):* A curated combination from Cleveland, Hungary, Switzerland, and Long Beach VA, focusing on features like exercise-induced angina.
5) *1,190-Record Dataset (Global):* Combines datasets from Cleveland, Hungary, and other locations, emphasizing 11 clinical features.
6) *4,240-Record Dataset (Framingham, USA):* A key focus due to its clinically relevant features (e.g., cholesterol, glucose) and imbalanced class distributions, mitigated by SMOTE and GANs.
7) *11,627-Record Dataset (USA):* Longitudinal data from the Framingham Heart Study covering cardiovascular risk factors.
8) *70,000-Record Dataset (Russia):* Focuses on cardiovascular disease indicators like alcohol intake and smoking.
9) *400,000-Record Dataset (USA):* Derived from the CDC BRFSS dataset, encompassing diverse features like physical activity levels and diabetes status.

These datasets provided a solid foundation for assessing generalizability and scalability across varying complexities.

### D. Research Questions and Findings

Key research questions and findings include:

1) *How do traditional ML models compared to neural network-based models in terms of accuracy and ROC AUC for heart failure prediction?*

ML models like RF achieved 83% accuracy and 0.91 ROC AUC on nonlinear datasets but struggled with high-dimensional data. While DL models, such as CNNs and RNNs, excelled in pattern recognition, achieving a 0.85 ROC AUC but incurred higher computational costs.

2) *What are the most influential predictors of heart failure across different datasets? Key predictors include:*
   - Large Datasets (400,000, 70,000, and 11,627 records): Age, BMI, systolic and diastolic blood pressure, and cholesterol.
   - Medium-Sized Datasets (4,240 records): Age, sysBP, cholesterol, and glucose.
   - Small Datasets (303, 1,000, and 1,025 records): Symptom-specific features like chest pain (cp).
3) *Can a hybrid stacking model that combines traditional ML and DL techniques provide superior predictive performance compared to single models?*

The hybrid stacking model combining RF, GBM, CNN, and RNN achieved 82% accuracy and 0.90 ROC AUC on small datasets and 90% accuracy with 0.97 ROC AUC on medium datasets.

4) *How does the use of Generative AI, particularly GANs, in a stacking model improve performance compared to standalone models? Does it enhance generalizability and scalability across diverse healthcare settings?*

GANs enhanced class balancing and improved ROC AUC from 0.83 (SMOTE) to 0.95 on the 4,240-record dataset.

5) *How does the unique Stacking Generative AI model specifically contribute to advancements in the healthcare industry, particularly in predicting and managing heart failure?*

*a)* Improved accuracy, class balance, and generalizability.

*b)* Enhanced clinical utility with personalized predictions and early intervention capabilities.

### E. Core Techniques and Optimization Strategies

#### 1) Synthetic Minority Over-Sampling Technique (SMOTE)

Addressed class imbalance by interpolating new data points for minority classes. On a 1,000-record dataset, SMOTE-enhanced models achieved a 0.95 ROC AUC.

#### 2) Grid Search Cross-Validation (Grid Search CV)

Optimized hyperparameters like RF's *n_estimators* (30) and *max_depth* (3) to improve model accuracy and AUC.

#### 3) Generative Adversarial Networks (GANs)

GANs generated synthetic samples by training a generator to create realistic data and a discriminator to validate its quality. This dual-network structure enhanced robustness and generalization in HF prediction.

*Complementary Role of SMOTE and GANs*

While SMOTE generates synthetic samples efficiently for traditional ML models, GANs produce realistic, high-quality data for complex, imbalanced datasets. Together with Grid Search Cross Validation (CV), these techniques enhance the model's performance, achieving superior accuracy and recall (Chawla et al., 2002; Goodfellow et al., 2014).

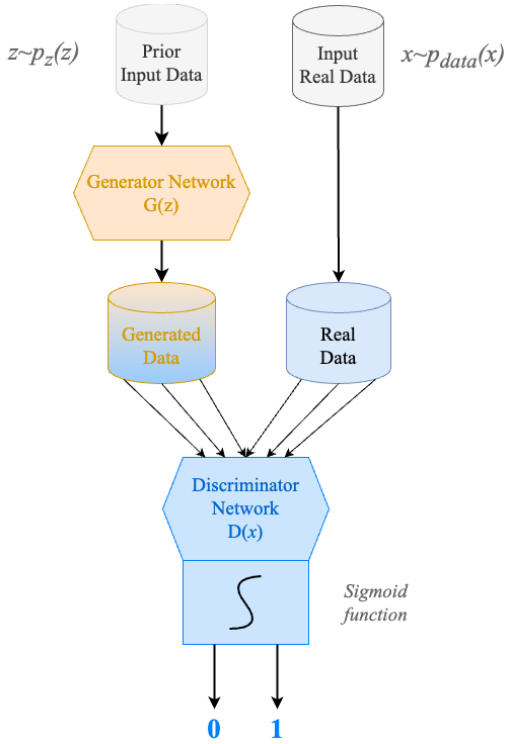
### F. Model Design and Implementation (Figure 2)

The Stacking Generative AI model combines traditional ML techniques, deep learning (DL) architectures, and synthetic data generated by Generative Adversarial Networks (GANs) to enhance heart failure prediction. Data preparation included cleaning, imputation using K-Nearest Neighbors (KNN) for numerical features and mode imputation for categorical features, and normalization through standard scaling to improve model convergence.

The GAN framework (Figure 1) integrates a generator and discriminator network. The generator, a feedforward neural

network, uses a latent space vector sampled from a Gaussian distribution to generate synthetic data. Its hidden layers are activated by ReLU, and the output layer utilizes Tanh to align synthetic data with normalized feature ranges. The discriminator, a binary classifier, includes hidden layers with LeakyReLU activation and a Sigmoid-activated output layer to distinguish between real and synthetic data. Both networks are trained using the Binary Cross-Entropy (BCE) loss function and the Adam optimizer at a learning rate of 0.00005.

*Figure 1- Architecture of the GAN network*



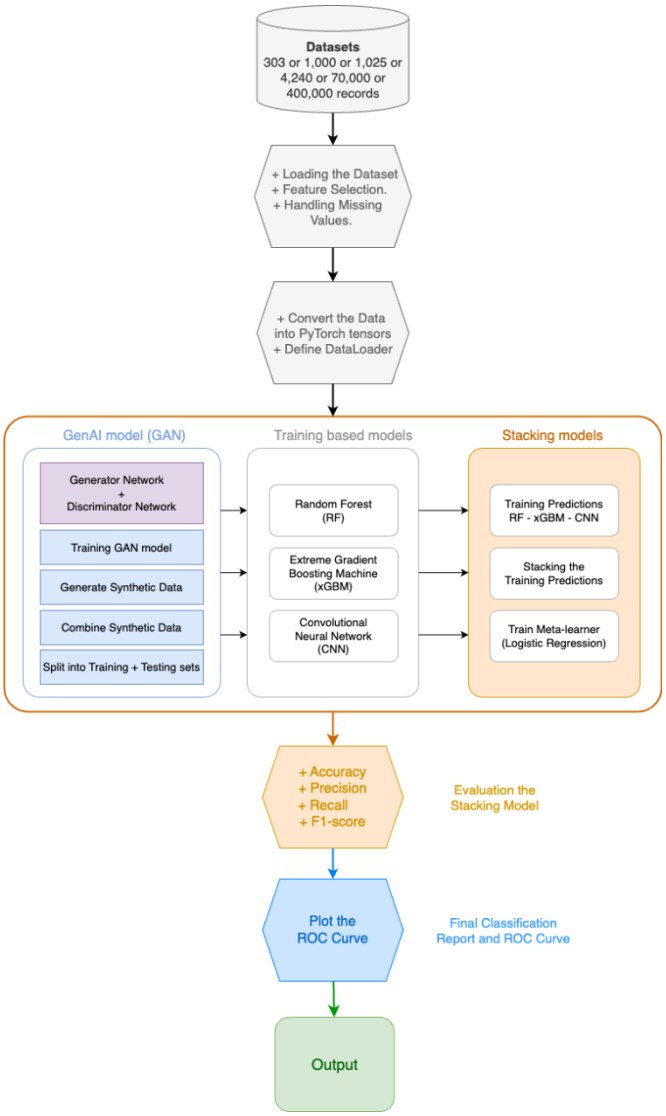

During GAN training, the generator aims to create synthetic samples that the discriminator misclassifies as real. The training process alternates updates between the two networks, using techniques like batch normalization, noise injection, and dropout to stabilize training and prevent mode collapse. The generated synthetic records are inverse-transformed to match the original feature space and validated by the discriminator for quality before integration into downstream tasks.

The stacking model incorporates base models, including Random Forest (RF), Extreme Gradient Boosting (xGBM), and Convolutional Neural Networks (CNNs). RF uses 100 trees and a maximum depth of 10, while xGBM employs 200 estimators and a learning rate of 0.05. The CNN architecture features Conv1D layers with MaxPooling and Dropout to prevent overfitting. Logistic Regression serves as the meta-learner, combining predictions from the base models. Its regularization parameter (C = 0.01) was fine-tuned to optimize the balance between bias and variance, ensuring robust final predictions.

### *G. Evaluation Measurement and Validation*

The Stacking Generative AI model underwent comprehensive evaluation to ensure robustness, generalizability, and real-world applicability. Statistical validation and cross-validation techniques assessed its performance across diverse datasets.

*Figure 2- Architecture of the Stacking Generative AI model*




Statistical validation included ANOVA analysis, revealing a significant improvement in performance ($p < 2.26e\text{-}276$), demonstrating the efficacy of GAN-generated data in addressing class imbalance. Mixed-effects modeling, as outlined by Javadi et al. (2023), accounted for variability across datasets, such as demographic differences between populations in Pakistan and the United States. Bias analysis showed potential overfitting in synthetic subsets, emphasizing the need for balancing real and synthetic data.

Cross-validation ensured model consistency and reliability. Using K-fold methods, including 5- and 10-fold cross-validation, the model achieved a mean accuracy of 99.4%. Learning curves identified areas of overfitting or underfitting during training and validation. Hyperparameter tuning through grid search optimized model parameters, such as the RF model's

30 estimators and maximum depth of 3, and the meta-learner's regularization strength (C = 0.01), further improving accuracy.

Comprehensive validation minimized overfitting and ensured generalization. Metrics such as sensitivity, specificity, and ROC AUC consistently demonstrated the model's ability to balance false positives and negatives. This rigorous evaluation confirms the Stacking Generative AI model's suitability for real-world deployment in heart failure prediction, offering high reliability and adaptability across diverse clinical scenarios.

The proposed Stacking Generative AI model underwent rigorous evaluation to ensure its robustness, generalizability, and reliability in real-world applications. Statistical validation and cross-validation techniques were employed to measure the model's performance across various datasets.

## IV. Results

### A. *Performance Comparison between Traditional Models and Neural Network Models: How do traditional ML models compare to neural network-based models in terms of accuracy and ROC AUC for heart failure prediction?*

The study compared traditional ML models, including LR, SVM, RF, GBM, and xGBM, with neural network models like CNN and GRU-based models for predicting heart failure. The performance metrics evaluated included accuracy and ROC AUC across datasets of varying sizes.

#### 1) *Performance on Small and Medium Datasets*

On smaller datasets (e.g., 303 records), RF performed best with 83% accuracy and a 0.91 ROC AUC, surpassing GBM (79% accuracy, 0.87 ROC AUC) and xGBM (80% accuracy, 0.86 ROC AUC). CNNs delivered comparable results with 82% accuracy and 0.85 ROC AUC. As dataset sizes increased to 1,000 and 1,025 records, RF and xGBM maintained strong results, achieving up to 93% accuracy and 0.98 ROC AUC. CNN's performance slightly declined (79% accuracy, 0.85 ROC AUC), while GRU-based models showed robust results with 84% accuracy and 0.92 ROC AUC, (Table 1).

*Table 1 – Small dataset's performances on ML and DL models*

| Dataset | Performance | Model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LR | SVM | RF | GBM | XGB | CNN | GRU w/ Attention | CNN w/ GRU |
| 303 | Accuracy | 79 | 85 | 83 | 79 | 80 | 82 | 80 | 80 |
| | ROC AUC | 85 | 86 | 91 | 87 | 86 | 85 | 84 | 87 |
| 1,000 | Accuracy | 80 | 85 | 90 | 88 | 88 | 79 | 77 | 78 |
| | ROC AUC | 86 | 92 | 94 | 94 | 95 | 85 | 84 | 84 |
| 1,025 | Accuracy | 82 | 81 | 91 | 91 | 93 | 82 | 80 | 84 |
| | ROC AUC | 91 | 91 | 95 | 97 | 98 | 93 | 86 | 92 |

#### 2) *Performance on Large Datasets*

Neural networks, particularly CNNs, excelled at handling large datasets. On a 400,000-record dataset, CNN achieved 78% accuracy and 0.86 ROC AUC, outperforming GBM (77% accuracy, 0.85 ROC AUC). RF delivered strong results with 90% accuracy and 0.96 ROC AUC, (Table 2).

*Table 2 – Large dataset's performances on ML and DL models*

| Dataset | Performance | Model | | | | | | | | Proposed |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LR | RF | GBM | XGB | CNN | GRU w/ Attention | CNN w/ GRU | Stacking ML+DL | Stacking Gen AI |
| 400,000 | Accuracy | 77 | 90 | 77 | 80 | 78 | 79 | 80 | 90 | 96 |
| | ROC AUC | 84 | 96 | 85 | 88 | 86 | 87 | 88 | 96 | 99 |

#### 3) *Comparative Analysis with Related Studies*

Compared to prior research, such as Dumlao, J. (n.d.), where RF achieved 94% accuracy, the Stacking Generative AI model demonstrated superior results on large datasets, achieving 96% accuracy and 0.99 ROC AUC. Similarly, when benchmarked against Khan, H. et al. (2024), whose models (EnsCVDD-Net and BlCVDD-Net) reported accuracies of 88% and 91%, and ROC AUCs of 0.88 and 0.91, the Stacking Generative AI model consistently outperformed, highlighting its robustness and precision in predicting heart failure, (Table 3).

*Table 3 – Large dataset's performances on ML and DL models*

| Dataset | Performance | Proposed Model | Compared Article | Compared Article |
|---|---|---|---|---|
| | | Stacking Generative AI | EnsCVDD-Net | BICVDD-Net |
| 400,000 | Accuracy | 96 | 88 | 91 |
| | ROC AUC | 99 | 88 | 91 |

### B. *What are the most influential predictors of heart failure across different datasets, and how do they affect overall model performance?*

Identifying key predictors enhances the accuracy and interpretability of heart failure models. Using RF's feature importance analysis, the study identified critical variables across datasets of different sizes:

#### 1) *Predictors in Large Datasets (Figure 3 and 4)*

In the 70,000-record dataset, key predictors included age, systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), and cholesterol, contributing to 74% accuracy and a 0.81 ROC AUC. The 400,000-record dataset highlighted, angina, BMI, and general health as top features, with the Stacking Generative AI model achieving 96% accuracy and a 0.99 ROC AUC.

*Figure 3 –Influential Predictors – dataset of 70,000 records*

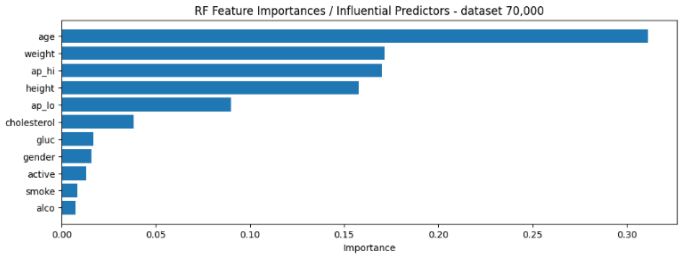
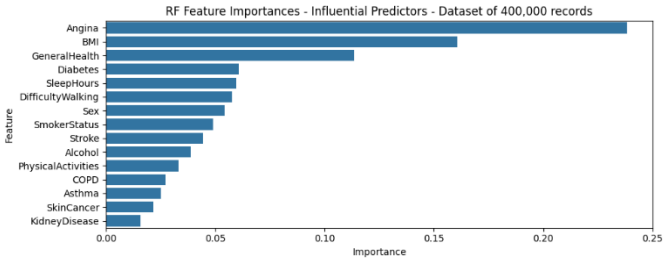


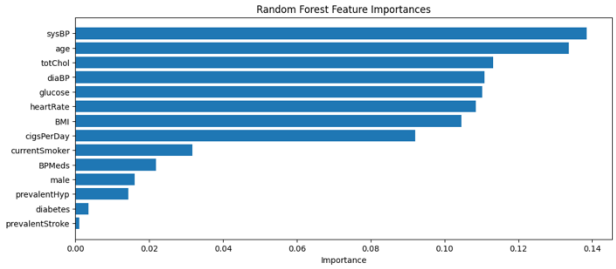

*Figure 4 –Influential Predictors – dataset of 400,000 records*



*2) Predictors in Medium-Sized Datasets (Figure 5)*

For the 4,240-record dataset, age, sysBP, and cholesterol were the strongest predictors, resulting in 92% accuracy and a 0.96 ROC AUC.

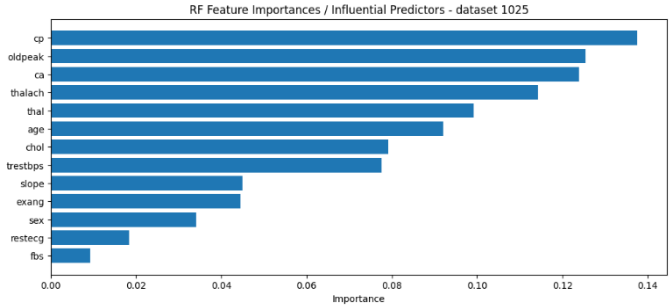
*Figure 5 – Influential Predictors – dataset of 4,240 records*




*3) Predictors in Small Datasets (Figure 6)*

In the 1,025-record dataset, chest pain (cp), oldpeak, and the number of major vessels (ca) were key predictors, achieving 95% accuracy and a 0.999 ROC AUC. While 303-record dataset identified heart rate attained (thalachh), chest pain (cp), and the number of major vessels (caa) as critical features, resulting in 95% accuracy and a 0.99 ROC AUC. For the 1,000-record dataset, the slope of the ST segment, chest pain (cp), and resting blood pressure were key contributors, yielding 98% accuracy and a 0.999 ROC AUC.

*Figure 6 – Influential Predictors – dataset of 1,025 records*




*4) Key Insights and Implications*

Blood pressure, chest pain, cholesterol levels, and age consistently emerged as the most critical predictors across datasets, aligning with established clinical risk factors for heart failure. These features improve the interpretability and clinical relevance of predictive models. By leveraging these predictors, the Stacking Generative AI model achieved superior accuracy and ROC AUC values, demonstrating the importance of systematic feature identification in advancing predictive healthcare.

*C. Can a hybrid stacking model that combines traditional ML and DL techniques provide superior predictive performance compared to single models?*

The study introduced a hybrid stacking model that combines traditional ML methods, such as RF and GBM, with advanced DL models like CNN and RNN. This approach leverages the strengths of ML and DL to improve predictive performance and scalability.

*1) Performance on Datasets of Varying Sizes*

On a small dataset (303 records), the hybrid stacking model achieved 82% accuracy and a ROC AUC of 0.90, outperforming standalone ML models like LR and SVM. The Stacking Generative AI model achieved a ROC AUC of 0.99, significantly exceeding RF (0.91) and SVM (0.86).

On a medium dataset (4,240 records), the stacking model reached 90% accuracy and a ROC AUC of 0.97, outperforming standalone CNN and RNN models. This highlights the hybrid model's ability to harness the strengths of both ML and DL techniques.

On a large dataset (400,000 records), the hybrid stacking model achieved 90% accuracy and a ROC AUC of 0.96, outperforming standalone models like CNN (78% accuracy, ROC AUC 0.86) and GBM (77% accuracy, ROC AUC 0.85). This demonstrates the scalability and robustness of the hybrid model.

*2) Comparative Analysis with Existing Models*

Compared to alternative hybrid approaches like Decision Tree with AdaBoost by Sk K. B. et al. (2023), which achieved 97.43% accuracy, the Stacking Generative AI model delivered competitive performance. On the Framingham dataset (4,240 records), the proposed model achieved 92% accuracy and a ROC AUC of 0.96, surpassing Mienye et al.'s CART-based ensemble (91% accuracy). And with the smallest dataset (303 records), the proposed model's ROC AUC of 0.99 exceeded the range reported by Rimal et al. (2024) (ROC AUC 0.85 to 0.95).

The hybrid stacking model outperforms standalone ML and DL models in both accuracy and ROC AUC. Its consistent superiority across datasets underscores its potential to advance predictive analytics in healthcare.

*D. How does the use of Generative AI, particularly GANs, in a stacking model improve performance compared to standalone models? Does it enhance generalizability and scalability across diverse healthcare settings?*

Generative AI, particularly GANs, enhances predictive performance by addressing class imbalance through synthetic data generation. This improves model training, reduces bias, and enhances scalability across diverse healthcare datasets.

*1) Performance Improvements*

On a small dataset (303 records), the Stacking Generative AI model achieved 95% accuracy and a ROC AUC of 0.99, outperforming standalone models like RF (83% accuracy, ROC AUC 0.91). Additionally, on a large dataset (400,000 records), the Stacking Generative AI model achieved 96% accuracy and a ROC AUC of 0.99, significantly exceeding CNN's 78% accuracy and ROC AUC of 0.86. This highlights the ability of GANs to improve performance even in large, complex datasets.

*2) Generalizability and Scalability*

The model maintained robust performance across datasets of varying sizes. On a 1,000-record dataset, it achieved 98% accuracy and a ROC AUC of 0.999, outperforming standalone CNN (79% accuracy) and RF (90% accuracy).

*3) Real-World Utility*

By addressing data imbalances and capturing intricate patterns, the Stacking Generative AI model supports scalable and predictive modeling. Its superior performance demonstrates

its potential for advancing clinical decision-making and improving real-world healthcare outcomes.

Generative AI, particularly GANs, plays an innovative role in predictive modeling by addressing key challenges like class imbalance and complex data patterns.

### *E. How does the unique Stacking Generative AI model specifically contribute to advancements in the healthcare industry, particularly in predicting and managing heart failure?*

The Stacking Generative AI model significantly advances HF prediction and management in the healthcare sector. By integrating traditional ML models such as RF, GBM, and xGBM with neural network algorithms like CNNs and RNNs, along with GANs, this proposed model addresses critical challenges such as class imbalance, scalability, and predictive accuracy.

#### *1) Class Imbalance Resolution*

GANs augment minority class data, improving recall and F1-scores. On a 303-record dataset, the model achieved 95% accuracy and a ROC AUC of 0.99, outperforming RF (83% accuracy, ROC AUC 0.91) and CNN (82% accuracy, ROC AUC 0.85) (Table 4).

#### *2) Scalability and Robustness*

The model scales effectively across datasets. On a 400,000-record dataset, it achieved 96% accuracy and a ROC AUC of 0.99, demonstrating its reliability for large-scale clinical applications and diverse patient populations.

#### *3) Clinical Utility*

By estimating key predictors like systolic blood pressure, cholesterol, and glucose, the model aids in early detection, risk stratification, and personalized treatment planning. Its accuracy make it a valuable tool for clinicians, optimizing resources and improving patient outcomes.

*Table 4 – Results across 12 models with evaluation on 9 datasets ranging from 299 – 400,000 records.*

| Dataset | Performance | Model | | | | | | | | | | Proposed | Current | Source |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LR | SVM | RF | GBM | XGB | CNN | GRU w/ Attention | CNN w/ GRU | Stacking ML+DL | Gen AI | Stacking Gen AI | Research Literature | Reference |
| **299** | Accuracy | 79 | 80 | 83 | 82 | 82 | 83 | 83 | 73 | 83 | 93 | 93 | 74 | Chico & Jurman (2020) |
| | ROC AUC | 86 | 88 | 92 | 87 | 90 | 87 | 84 | 83 | 91 | 98 | 98 | 80 | *(RF)* |
| **303** | Accuracy | 79 | 85 | 83 | 79 | 80 | 82 | 80 | 80 | 82 | 95 | 95 | 93 | Rimal, Y. et al. (2024) |
| | ROC AUC | 85 | 86 | 91 | 87 | 86 | 85 | 84 | 87 | 90 | 99 | 99 | 90 | *(RF & SVM)* |
| **1,000** | Accuracy | 80 | 85 | 90 | 88 | 88 | 79 | 77 | 78 | 94 | 98 | 98 | 97 | Dumlao, J. (n.d.) |
| | ROC AUC | 86 | 92 | 94 | 94 | 95 | 85 | 84 | 84 | 98 | 99 | 99.9 | NA | *(RF)* |
| **1,025** | Accuracy | 82 | 81 | 91 | 91 | 93 | 82 | 80 | 84 | 95 | 95 | 98 | 83 | Elfar H. (n.d.) |
| | ROC AUC | 91 | 91 | 95 | 97 | 98 | 93 | 86 | 92 | 98 | 99 | 99.9 | 93 | *(RF)* |
| **1,190** | Accuracy | 86 | 88 | 92 | 90 | 92 | 88 | 86 | 86 | 92 | 96 | 98 | 90 | Liu et al. (2022) |
| | ROC AUC | 93 | 95 | 96 | 95 | 96 | 94 | 93 | 93 | 96 | 99 | 99.9 | 95 | *(ML stacking)* |
| **4,240** | Accuracy | 65 | 67 | 71 | 81 | 86 | 70 | 63 | 67 | 90 | 93 | 92 | 91 | Mienye et al. (2020) |
| | ROC AUC | 74 | 74 | 79 | 89 | 93 | 77 | 70 | 72 | 97 | 96 | 96 | NA | *(CART)* |
| **11,627** | Accuracy | 71 | 78 | 84 | 79 | 83 | 74 | 74 | 73 | 85 | 91 | 91 | 97 * | Sk K. B. et al (2023) |
| | ROC AUC | 79 | 85 | 92 | 88 | 92 | 83 | 82 | 84 | 93 | 95 | 95 | NA | *(DT+AdaBoost)* |
| **70,000** | Accuracy | 72 | 73 | 73 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 74 | 72 | Jain, S. (n.d.) |
| | ROC AUC | 79 | 79 | 79 | 80 | 81 | 80 | 81 | 80 | 81 | 80 | 81 | NA | *(ML stacking)* |
| **400,000** | Accuracy | 77 | NA | 90 | 77 | 80 | 78 | 79 | 80 | 90 | 95 | 96 | 91 | Khan, H. et al. (2024) |
| | ROC AUC | 84 | NA | 96 | 85 | 88 | 86 | 87 | 88 | 96 | 98 | 99 | 91 | *(BICVDD-Net)* |

#### *4) Summary of Proposed Model Contributions*

Combining predictive accuracy, scalability, and generalizability, the Stacking Generative AI model represents an innovative tool for HF prediction and management. It addresses challenges in healthcare data, advancing early detection and personalized care, and improving clinical decision-making (Table 4, Figure 7, 8).

## V. Conclusion

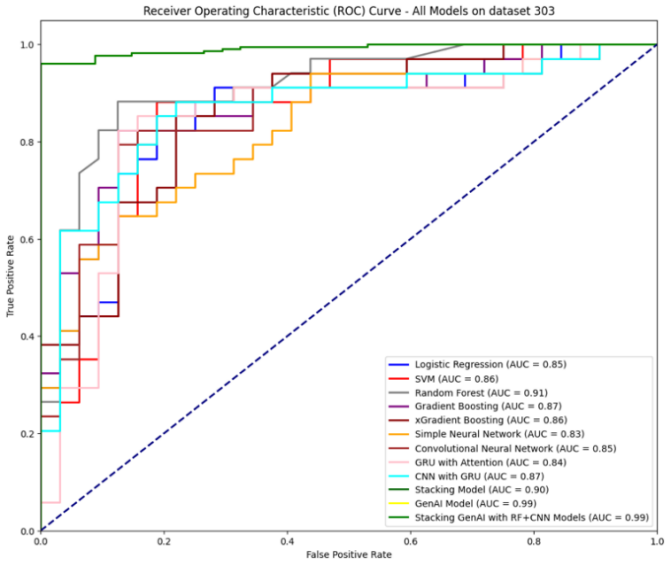
This study highlights the effectiveness of the Stacking Generative AI model as an innovative hybrid solution for heart failure prediction. By integrating Generative AI with traditional machine learning models (RF, GBM, xGBM) and deep learning architectures (CNNs, RNNs), the model addresses key challenges in healthcare predictive modeling, including class imbalance, scalability, and predictive accuracy. Its ability to synthesize balanced datasets and adapt to varying complexities makes it a robust tool for clinical applications.

### *A. Summary of Findings*

The Stacking Generative AI model consistently outperformed standalone ML and DL models across datasets of varying sizes, from 299 to 400,000 records. Notably, it achieved an outstanding ROC AUC of 0.999 on a 1,000-record dataset, surpassing standalone xGBM (0.94) and CNN (0.85). On the largest dataset of 400,000 records, it maintained performance with 96% accuracy and an ROC AUC of 0.99, demonstrating its scalability and applicability across clinical settings.

The hybrid structure of the model leverages the strengths of ML and DL techniques. While ML models excel at structured data analysis, DL models identify complex patterns in data. The integration of GANs enhances the model's ability to balance minority classes, improving recall and F1-scores—features that are critical for identifying high-risk cases in healthcare, where class imbalances are prevalent.

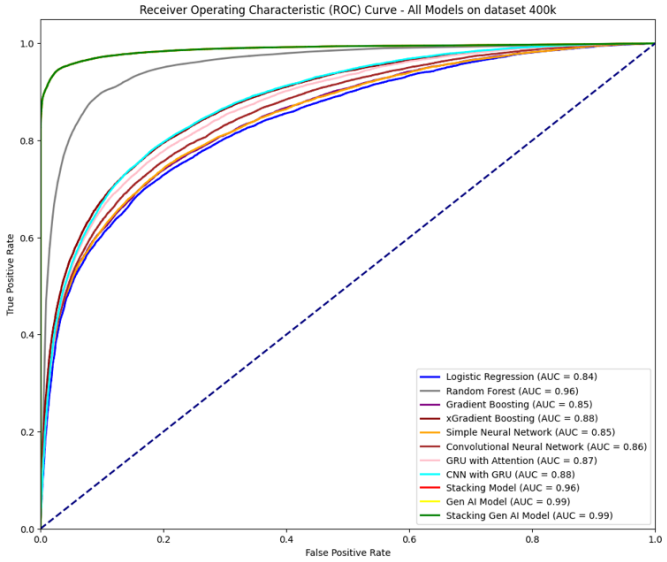
*Figure 7- ROC AUC performance on dataset of 303 records*




## B. *Comparison with Existing Literature*

Compared to prior studies, the Stacking Generative AI model demonstrates clear advancements. Singh et al. (2024) reported an ROC AUC of 0.89 using xGBM, while the proposed model achieved 0.99 on similar datasets. Similarly, the model outperformed RF-based approaches, such as those in Chicco et al. (2022), which achieved an ROC AUC of 0.85. These comparisons affirm the utility of combining GANs with ML and DL techniques within a stacking framework, resulting in superior predictive accuracy and scalability.

*Figure 8- ROC AUC performance on dataset of 400,000 records*




## C. *Clinical Implications*

The Stacking Generative AI model has significant potential for clinical implementation. Its ability to handle imbalanced datasets while delivering high predictive accuracy makes it a reliable decision-support tool for early diagnosis and personalized treatment planning. By identifying actionable predictors such as systolic blood pressure, BMI, cholesterol, and glucose levels, the model provides clinicians with valuable insights, enabling better resource allocation and improved patient care.

The model's adaptability across diverse datasets further underscores its utility. Its consistent performance across datasets of varying sizes, from 299 to 400,000 records, highlights its robustness and scalability. Although its complexity may reduce interpretability compared to simpler models like logistic regression, the trade-off is justified by its superior predictive power.

## D. *Limitations and Future Research*

Despite its achievements, the model faces certain limitations. Discrepancies in the fidelity of GAN-generated data, particularly for binary and categorical variables, require further investigation. Future research will focus on improving GAN architectures, such as transformer-based GANs, to enhance the quality of synthetic data. Additionally, incorporating feature-specific loss functions could align synthetic data more closely with original datasets, particularly for critical clinical variables. Expanding the validation of the model to include larger and more diverse datasets will further enhance its generalizability and clinical relevance.

## E. *Conclusion*

The proposed Stacking Generative AI model represents a significant advancement in heart disease prediction. By combining ML, DL, and GANs, the model achieves accuracy, scalability, and generalizability, outperforming existing models in both existing literatures and real-world scenarios. The model's application in healthcare systems has the potential to transform heart failure prediction and management by enabling early diagnosis, personalized care, and data-driven decision-making. Furthermore, the study designed and developed a web application to demonstrates (https://cvdstack.streamlit.app) its practical utility, offering real-time risk assessment tools for clinicians and patients. By bridging the gap between academic research and clinical practice, this model paves the way for future advancements in predictive healthcare analytics.

# VI. Discussion and Future Works

## A. *Discussion*

This study assessed the Stacking Generative AI model's performance in heart failure prediction, focusing on computational efficiency, memory usage, and clinical integration. With an inference time of 0.0095 seconds per prediction and memory usage of 1278.99 MB, the model proved suitable for real-time applications. Cloud-based deployment was identified as a cost-effective solution for scalability. The model's integration into clinical workflows, such as EHRs, was emphasized, showcasing its ability to augment predictive systems while safeguarding patient privacy.

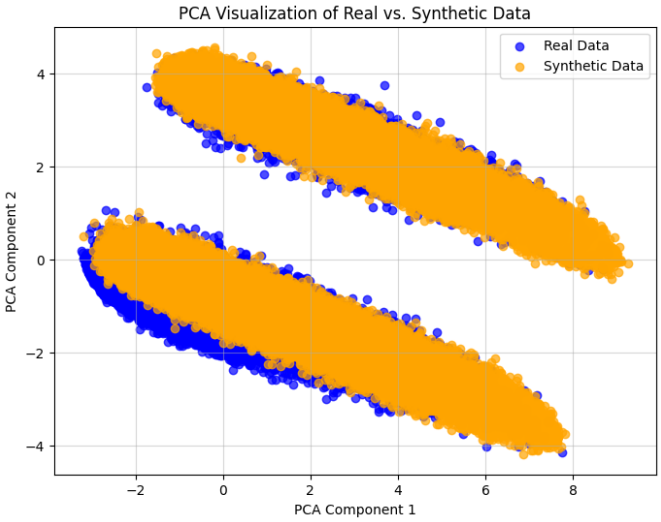
Despite strong metrics (accuracy: 92%, ROC AUC: 0.96 on the 4,240-record dataset), discrepancies between GAN-generated synthetic data and original data revealed opportunities for improvement. Refinements, including feature-specific loss functions and advanced GAN architectures, such as Wasserstein GANs, are proposed to enhance data fidelity and mitigate discrepancies. Validation techniques like SHAP analysis and

feature-wise tests will further evaluate the impact of these improvements.

*Comparison Between GAN-Generated Synthetic Data and Original Data*

The analysis of the 400,000-record dataset reveals several key findings about the relationship between GAN-generated synthetic data and original data. The PCA visualization (Figure 9) highlights significant overlap between real and synthetic data distributions, reflecting the GAN's ability to replicate patterns while introducing necessary diversity. However, the broader spread of synthetic data indicates room for refinement in aligning synthetic samples more closely with the original data.

*Figure 9 – PCA Visualization of Original vs. Synthetic Data for the 400,000-Record Dataset*




These results emphasize the dual role of synthetic data in addressing class imbalances and introducing biases. For the 400,000-record dataset, synthetic data improved overall metrics (ROC AUC: 0.99), demonstrating its efficacy in augmenting minority class samples. However, when evaluated on the original subset alone, the model's performance declined significantly (ROC AUC: 0.4586), underscoring the importance of ensuring alignment between synthetic and original data distributions.

The imbalance in the original dataset exacerbated these challenges, as underrepresentation of minority cases hindered generalization. Future work should focus on refining GAN architectures and preprocessing strategies to address these biases and further improve model robustness and equity in predictions for large-scale datasets.

*Performance Variability Across Datasets*

Model performance varied across datasets due to differences in size, class distribution, and feature diversity. Smaller datasets (e.g., 299 and 303 records) exhibited high sensitivity (≥0.97) but lower specificity (0.65–0.78), reflecting a focus on detecting heart failure cases. Larger datasets (e.g., 11,627 and 400,000 records) showed balanced sensitivity (0.83–0.95) and specificity (0.97–0.99), demonstrating improved generalization. The 70,000-record dataset's lower sensitivity (0.67) and F1-score (0.74) highlighted challenges with class imbalance and noise. Advanced resampling techniques and hyperparameter optimization are recommended to address these issues.

*Table 5 – Performance Metrics (Sensitivity, Specificity, and F1-Score) for Stacking Gen AI Model Across Datasets*

| **Dataset** (Records) | **Sensitivity** (Recall for Class 1) | **Specificity** (Recall for Class 0) | **F1 Score** (Macro Avg) | **ROC AUC** |
|---|---|---|---|---|
| 299 | 0.97 | 0.78 | 0.88 | 0.98 |
| 303 | 0.99 | 0.65 | 0.86 | 0.99 |
| 1,000 | 0.99 | 0.95 | 0.97 | 1.00 |
| 1,025 | 1.00 | 0.92 | 0.97 | 1.00 |
| 1,190 | 0.99 | 0.96 | 0.97 | 1.00 |
| 4,240 | 0.88 | 0.97 | 0.92 | 0.96 |
| 11,627 | 0.83 | 0.99 | 0.91 | 0.95 |
| 70,000 | 0.67 | 0.81 | 0.74 | 0.81 |
| 400,000 | 0.95 | 0.97 | 0.96 | 0.99 |

*Evaluation Metrics Across Combined, Original, and Synthetic Datasets*

The combined dataset of 1,241 records achieved strong metrics (accuracy: 88%, ROC AUC: 0.9311). However, the original subset's performance dropped (accuracy: 42%, ROC AUC: 0.4586), reflecting challenges with imbalanced data. Conversely, the synthetic subset achieved perfect metrics, underscoring overfitting risks. Balancing original and synthetic data is critical for ensuring robustness and mitigating biases, particularly in clinical scenarios.

*Table 6 – Evaluation Metrics Across Datasets*

| **Dataset** | **Record** | **Accuracy** | **Sensitivity** | **Specificity** | **F1 Score** | **ROC AUC** |
|---|---|---|---|---|---|---|
| Combined (Original + Synthetic) | 1,241 | 88% | 94% | 84% | 87% | 0.931 |
| Original Subset | 732 | 42% | 53% | 39% | 24% | 0.458 |
| Synthetic Subset | 509 | 100% | 100% | N/A | 100% | N/A |

## B. *Future Works*

Future research will address the limitations and explore avenues for enhancing the model's performance and applicability:

1) *Evaluation on External Datasets:* To improve generalizability, future studies will include datasets from diverse demographics, including underrepresented populations from Africa, South America, and East Asia. Subgroup analyses will identify potential biases and ensure equitable model performance.
2) *Exploring Advanced Models:* Transformer-based architectures and reinforcement learning will be incorporated to improve sequential tasks and predictions. Integration of large language models with EHRs could further enhance predictive accuracy.

3) *Expanding Applications:* Adapting the model for other medical conditions, such as diabetes and chronic kidney disease, will demonstrate broader utility in healthcare.
4) *Enhancing Interpretability:* Developing counterfactual explanations and visualization tools will improve clinical usability, ensuring the model's predictions are interpretable and actionable.
5) *Real-World Validation:* Clinical trials will refine the model based on healthcare practitioners' feedback, focusing on usability and practical challenges.
6) *Improving Efficiency:* Techniques like model pruning, quantization, and edge computing will optimize performance for resource-constrained environments.

The web/mobile application developed for this study (accessible at [https://cvdstack.streamlit.app/]) demonstrates practical deployment potential. The application allows clinicians and patients to input clinical data, such as age, cholesterol, and blood pressure, to obtain real-time heart failure risk predictions. Future usability testing will evaluate task completion times, user satisfaction, and scalability under high workloads, refining the interface for clinical integration.

### C. Ethical Considerations

The use of GAN-generated synthetic data raises ethical concerns. While synthetic data mitigates class imbalance and preserves patient privacy, it may propagate biases from the original dataset, particularly for underrepresented groups. Fairness testing and privacy-preserving techniques, such as differential privacy, will ensure ethical standards in data generation and model predictions.

### D. Summary

The Stacking Generative AI model represents a significant advancement in predictive healthcare. By combining ML, DL, and GANs, the model addresses challenges such as class imbalance, scalability, and accuracy, making it a robust tool for heart failure prediction. Future efforts will focus on refining GAN architectures, expanding its applications, and validating its real-world utility, ensuring maximum impact on clinical care.

## REFERENCES


[1] Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC Medical Informatics and Decision Making, 20 (2020): 1-16.

[2] Singh, M.S., Thongam, K., Choudhary, P., Bhagat, P.K. "An Integrated Machine Learning Approach for Congestive Heart Failure Prediction." Diagnostics, 14(7):736, 2024.

[3] Rimal, Y., & Sharma, N. "Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy." Multimedia Tools and Applications, 2024.

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

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. "Generative adversarial networks." Communications of the ACM, 63(11):139-144, 2014.

[6] Ng, A. Y., & Jordan, M. I. "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes." NIPS, 2001.

[7] Khan, H., et al. "EnsCVDD-Net and BlCVDD-Net: Models for heart disease prediction." Journal of Cardiovascular Studies, 2024.

[8] Yu, F., et al. "Feature-enhanced loss functions in GAN frameworks for coronary artery disease prediction." KORA cohort study, 2024.

[9] Mienye, I. D., Sun, Y., & Wang, Z. "Hybrid and ensemble models for heart disease prediction." Expert Systems with Applications, 2020.

[10] Wankhede, D., et al. "DL techniques and swarm algorithms for medical datasets." Journal of Computational Intelligence and Healthcare, 2022.

[11] Arooj, S., et al. "Neural network-based heart disease prediction." International Journal of Advanced Computer Science and Applications, 2022.

[12] Frid-Adar, M., et al. "GANs for medical imaging and predictive analysis." 2018 International Conference on AI and Healthcare, 2018.

[13] Yi, X., et al. "Combining GANs with traditional ML techniques in healthcare datasets." Medical Data Symposium, 2019.

[14] Bhagawati, M., & Paul, S. (2024, March). Generative Adversarial Network-based Deep Learning Framework for Cardiovascular Disease Risk Prediction. IEEE.

[15] Liu, J., Dong, X., Zhao, H., & Tian, Y. (2022). Predictive classifier for cardiovascular disease based on stacking model fusion. *Processes*, *10*(4), 749.

[16] Sk, K. B., Roja, D., Priya, S. S., Dalavi, L., Vellela, S. S., & Reddy, V. (2023, March). Coronary Heart Disease Prediction and Classification using Hybrid Machine Learning Algorithms. IEEE.

[17] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, 208-222.

[18] Tuli, Shreshth, et al. "HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments." Future Generation Computer Systems 104 (2020): 187-200.

[19] Mahmud, Istiak, et al. "Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel." Diagnostics 13.15 (2023): 2540.

[20] Hasan, Omar Shakir, and Ibrahim Ahmed Saleh. "DEVELOPMENT OF HEART ATTACK PREDICTION MODEL BASED ON ENSEMBLE LEARNING." Eastern-European Journal of Enterprise Technologies 112 (2021).

[21] Javadi, M., Sharma, R., Tsiamyrtzis, P. *et al.* Let UNet Play an Adversarial Game: Investigating the Effect of Adversarial Training in Enhancing Low-Resolution MRI. *J Digit Imaging. Inform. med.* (2024). https://doi.org/10.1007/s10278-024-01205-8

[22] M. Javadi, R. Sharma, P. Tsiamyrtzis, S. Shah, E. L. Leiss and N. V. Tsekos, "From Perception to Precision: Navigating Perceptual Loss in MRI Super-Resolution," *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, Dayton, OH, USA, 2023, pp. 57-61, doi: 10.1109/BIBE60311.2023.00017.

[23] Sharma, R., Tsiamyrtzis, P., Webb, A.G. et al. Learning to deep learning: statistics and a paradigm test in selecting a UNet architecture to enhance MRI. Magn Reson Mater Phy 37, 507–528 (2024). https://doi.org/10.1007/s10334-023-01127-6

[24] Sharma, R., Tsiamyrtzis, P., Webb, A. G., Seimenis, I., Loukas, C., Leiss, E., & Tsekos, N. V. (2022). A Deep Learning Approach to Upscaling "Low-Quality" MR Images: An In Silico Comparison Study Based on the UNet Framework. Applied Sciences, 12(22), 11758. https://doi.org/10.3390/app122211758

[25] Chen, H., et al. (2023). Hyperparameter tuning in healthcare models. International Journal of Data Science, 19(1), 9