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 seven datasets ranging from 303 to 400,000 records. Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalances, while a novel 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 novel 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) represents a growing public health concern due to its high morbidity and mortality rates. Early detection and timely intervention are essential for improving patient outcomes and reducing the burden on healthcare systems. Predictive models play a vital role in enabling timely and informed decision-making for clinicians and patients (Davis & Smith, 2023). In recent years, ML and DL techniques have shown significant promise in healthcare applications, particularly for predictive tasks (Breiman, 2001; LeCun et al., 2015).

Traditional ML models, including LR, RF, GBM, and xGBM, have demonstrated effectiveness in predictive applications. However, these models often fail to capture the nonlinear and temporal complexities of healthcare datasets. Conversely, neural networks, such as CNNs and RNNs, excel at identifying complex patterns of data but are constrained by computational demands and limited interpretability, making them less suitable for widespread clinical adoption (LeCun et al., 2015; Cho et al., 2014).

Hybrid models, particularly stacking approaches, address these limitations by combining the strengths of diverse algorithms. These models use a meta-learner to integrate predictions from base models, improving accuracy and generalizability (Sagi & Rokach, 2018). This study introduces a novel Stacking Generative AI (Gen AI) model that integrates Generative Adversarial Networks (GANs), RF, GBM, xGBM, and CNNs to predict HF (Goodfellow et al., 2014). GANs play a crucial role by generating synthetic data to address class imbalances, enhancing model performance on imbalanced datasets commonly found in healthcare (Frid-Adar et al., 2018; Yi et al., 2019).

The research aims to evaluate the performance of traditional ML models, DL models, standalone Generative AI, and the proposed Stacking Generative AI model across seven diverse datasets. Key research questions include: (a) How do traditional ML models compare to DL models like CNN and RNN? (b) What are the most influential predictors of HF, and how do these features influence model performance? (c) Can a hybrid stacking model combining ML, DL, and GANs outperform single models? (d) How does incorporating Generative AI improve stacking model performance? (e) What contributions can the Stacking Generative AI model make to healthcare, particularly in HF prediction and management?

Initial findings reveal that the proposed hybrid approach consistently outperforms traditional and standalone models. On these datasets of 1,000 and 1,025 records, the Stacking Generative AI model both achieved 98% accuracy and ROC AUC of 0.999. This demonstrates its capacity to address class imbalance, capture complex data patterns, and enhance clinical relevance. By integrating advanced AI techniques, this research underscores the novel potential of hybrid models in improving HF prediction and advancing personalized patient care.

II. Literature Review

A. Traditional ML Approaches in HF Prediction

Traditional ML models, such as RF, GBM, xGBM, and LR, have demonstrated robust performance in predicting HF. However, these models often face challenges in handling nonlinear relationships, class imbalances, and high-dimensional healthcare data. Chicco and Jurman (2020) identified RF as the best-performing model for HF survival prediction, achieving an accuracy of 74% and a ROC AUC of 0.80. Despite its effectiveness, the study's limited dataset and narrow feature scope restricted its generalizability.

Other studies leveraged larger datasets to improve performance. Singh et al. (2024) employed advanced preprocessing techniques, including feature selection with C4.5 and imputation with K-Nearest Neighbor (KNN), to train a Deep Neural Network (DNN) on a dataset of 5,888 records. This approach yielded a 95.3% accuracy and a ROC AUC of 0.97, although noisy and complex datasets remained a challenge. Similarly, Rimal et al. (2024) optimized RF using Bayesian optimization and genetic algorithms, achieving 89% accuracy. While effective, these methods lacked scalability and struggled with generalization.

Ensemble approaches offer promising solutions to these challenges. Hasan and Saleh (2021) applied stacking to the Framingham dataset (4,239 records), achieving a 96.69% accuracy and a ROC AUC of 0.98. However, the absence of integration with DL or Generative AI techniques limited its potential. In contrast, the proposed Stacking Generative AI model, achieved superior performance with 95% accuracy and a 99% ROC AUC by addressing scalability, class imbalance, and complex data patterns.

B. Neural Network-Based Approaches

Deep learning (DL) models have significantly advanced HF prediction by capturing complex data patterns that traditional ML models often miss. Mahmud et al. (2023) introduced a lightweight metamodel combining ML algorithms, achieving an 87% accuracy on a dataset of 920 records. Although efficient, this model was less capable of capturing intricate patterns compared to advanced DL methods.

Recurrent Neural Networks (RNNs), particularly with Gated Recurrent Units (GRUs), have been effective in temporal sequence modeling. Choi et al. (2017) achieved a ROC AUC of 0.883 using RNNs on electronic health record (EHR) data. However, the absence of hybrid or ensemble strategies limited broader applicability. Similarly, Arooj et al. (2022) employed a Deep Convolutional Neural Network (DCNN) on a 1,050-record dataset, achieving a 91.7% accuracy. While effective, the single-dataset approach restricted generalizability.

Recent advancements, such as transformers, also show promise. Sakthi et al. (2024) utilized transformers to identify heart anomalies, achieving an 88.6% accuracy. However, hybrid approaches combining DL with ML were not explored. Tuli et al. (2020) proposed HealthFog, an IoT-based framework integrating ensemble DL with fog computing, achieving a 91.2% accuracy and a ROC AUC of 0.94. Despite its scalability for real-time applications, reliance on device resources posed challenges for widespread adoption.

C. Hybrid and Stacking Models in HF Prediction

Hybrid models improve prediction accuracy by integrating multiple algorithms to compensate for individual weaknesses. Ali et al. (2020) developed a DL-based system combining wearable sensor data with electronic medical records (EMRs), achieving a 98.5% accuracy. However, relying solely on DL limited the model's generalizability.

Mienye et al. (2020) achieved 93% accuracy using ensemble ML models but excluded DL techniques essential for capturing complex data patterns. Wankhede et al. (2022) introduced a hybrid ensemble model combining DL with the Tunicate Swarm Algorithm, achieving 97.5% accuracy on the Cleveland dataset. Nonetheless, the model's small dataset and lack of ML integration hindered scalability. Liu et al. (2022) achieved ROC AUCs of 0.95 and 0.92 on two datasets using stacking with multiple classifiers but did not employ Generative AI methods to enhance performance.

D. Generative AI and GAN Frameworks in HF Prediction

Generative Adversarial Networks (GANs) have emerged as effective tools for addressing class imbalances and data complexity in HF prediction. Khan et al. (2024) integrated ML and DL with GANs to generate synthetic data, achieving a 96.1% accuracy and a ROC AUC of 0.927. Yu et al. (2024) proposed a GAN framework with a feature-enhanced loss function, achieving a 94.62% accuracy and a ROC AUC of 0.958 on the KORA cohort dataset. Similarly, Bhagawati and Paul (2024) applied GANs to coronary artery disease prediction, achieving a 93% accuracy and a ROC AUC of 0.953.

The Stacking Generative AI model builds on these advancements by synthesizing balanced datasets and enhancing minority class prediction. This results in superior performance, with 95% accuracy and a 99% ROC AUC across multiple datasets.

III. Methodology

The methodology introduces an approach for evaluating ML, DL, and standalone Generative AI (Gen AI) models for heart failure (HF) prediction. It focuses on the Stacking Generative AI model, which combines these techniques to address class imbalance and improve prediction accuracy.

A. Generative AI Models

1) Stacking Generative AI Model

This hybrid framework combines ML models (RF, GBM, xGBM) and DL architectures (CNNs, RNNs). Generative AI synthesizes balanced datasets to address class imbalance and boost generalizability. For small datasets, traditional ML techniques ensured reliable performance, while larger datasets leveraged CNNs and RNNs for handling complexity. This adaptability achieved 98% accuracy and a ROC AUC of 99.9% on 1,025-record dataset and 96% accuracy with a ROC AUC of 0.99 on 400,000-record dataset.

2) Standalone Generative AI Models

Standalone Generative AI models achieved robust results, enhancing predictions for minority classes. In particular, a standalone model reached a ROC AUC of 0.99 on a dataset with 4,240 records, showcasing the potential of synthetic data to mitigate data limitations and imbalances.

B. Overview of Methodology

The methodology involved preprocessing and analyzing HF datasets ranging from 303 to 400,000 records. Preprocessing steps included data cleaning, missing values, normalization, and balancing using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, and feature scaling. Z-score normalization ensured optimal model convergence.

SMOTE generated synthetic data for minority classes, enhancing the performance of both ML and DL models. GANs method further improved generalizability and robustness by creating high-quality synthetic data. Extensive hyperparameter tuning using Grid Search Cross-Validation optimized model performance. Metrics such as accuracy, ROC AUC, precision, recall, and F1-scores were used for evaluation, with the Stacking Generative AI model consistently demonstrating superior results.

C. Data Collection and Preprocessing

Seven datasets were selected to ensure diversity and relevance in capturing heart disease indicators. These datasets varied in size and complexity, providing a robust foundation for evaluating model performance:

1) *Cleveland Heart Disease Dataset:* 303 records with 14 features, historically achieving up to 85% accuracy.
2) *Indian Heart Disease Dataset:* 1,000 records, achieved accuracies up to 94% with decision trees.
3) *Combined Cleveland, Hungary, Switzerland, and Long Beach Dataset:* 1,025 records offering global population diversity, achieving up to 89% accuracy.
4) *Framingham Dataset:* 4,240 records for 10-year risk analysis, achieving 80% to 90% accuracy.
5) *Framingham Heart Study Dataset:* Longitudinal data of 11,627 records for cardiovascular trends, achieving up to 92% accuracy.
6) *Kaggle Dataset:* 70,000 records demonstrating model scalability, achieving up to 73% accuracy.
7) *BRFSS Dataset:* 400,000 records supporting public health research with up to 88% accuracy.

These datasets provided a comprehensive basis for testing generalizability, scalability, and robustness across various model architectures.

D. Research Questions and Core Findings

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

Traditional ML models, like RF and GBM, perform well on nonlinear datasets, achieving accuracies of 83% and 79%, with ROC AUCs of 0.91 and 0.88 respectively. However, they struggle with high-dimensional data. Neural networks, such as CNNs and RNNs, excel at capturing complex relationships, achieving ROC AUCs of 0.85 and 0.80 on smaller datasets. Still, their performance decreases with larger datasets, showing higher computational costs but better pattern recognition.

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 blood pressure (ap_hi), diastolic blood pressure (ap_lo), 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), this finding significantly improves model accuracy and interpretability, enhancing clinical relevance.

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

Yes, the hybrid stacking model combining RF, GBM, CNN, and RNN achieved superior performance. It recorded 82% accuracy with a 0.90 ROC AUC on a small dataset (303 records) and 90% accuracy with a 0.97 ROC AUC on a medium-sized dataset (4,240 records). This demonstrates the robustness and versatility of hybrid models in combining the strengths of ML and DL.

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?

Generative Adversarial Networks (GANs) improved performance significantly by generating high-quality synthetic data. On a 4,240-record dataset, SMOTE method achieved the ROC AUC of 0.83 while GAN achieved 0.95 in ROC AUC. GANs enhanced class balance, generalizability, and scalability across diverse datasets.

5) 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 integrates ML, DL, and Generative AI to:

a) Improve Accuracy: Achieving superior results by capturing complex data patterns.

b) Handle Class Imbalance: GANs generate high-quality synthetic samples for minority classes.

c) Enhance Generalizability: Adapting to diverse datasets and patient populations.

d) Support Personalized Care: Providing risk predictions for tailored treatments and early interventions.

e) Increase Clinical Utility: Offering accessible tools for clinicians and patients.

E. Core Techniques and Optimization Strategies

1) Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE addresses class imbalances by interpolating new data points between minority class instances (Chawla et al., 2002). As a result, on a 1,000-record dataset, SMOTE-enhanced xGBM and stacking models achieved ROC AUCs of 0.95 and 0.98, respectively. For larger datasets, SMOTE added synthetic samples (e.g., 219,152 for a 400,000-record dataset), improving recall and precision for minority classes with less overfitting.

Mathematically, SMOTE generates synthetic samples using:

$$x_{\text{new}} = x_{\text{minority}} + \lambda \cdot (x_{\text{neighbor}} - x_{\text{minority}})$$

where x_{minority} is a minority class instance, x_{neighbor} is one of its nearest neighbors, and λ is a random number between 0 and 1. This process creates a more diverse minority class dataset without simply duplicating existing instances, Chawla et al. (2002).

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

Grid Search CV optimized hyperparameters for base models (e.g., RF, xGBM, CNN) and the meta-learner in the Stacking Generative AI model. Key parameters, such as RF's n_estimators = 30 and max_depth = 3 were fine-tuned, ensuring optimal performance before integration into the meta-learner. This systematic tuning enhanced accuracy and AUC across datasets.

3) Generative Adversarial Networks (GANs)

GANs generate synthetic data by training a Generator Network to create realistic patient profiles. To illustrate, consider the case of latent inputs (G(z)) produce features like systolic blood pressure and cholesterol, enriching training sets and improving model robustness (Goodfellow et al., 2014).

The Discriminator Network acts as a binary classifier, distinguishing between real and synthetic data. Using LeakyReLU activations in hidden layers and Sigmoid in the output layer, it validates the quality of synthetic data, ensuring it closely resembles actual patient data. This validation improves model generalization, aiding early heart failure detection and preventive care (Radford et al., 2015).

Complementary Role of SMOTE and GANs

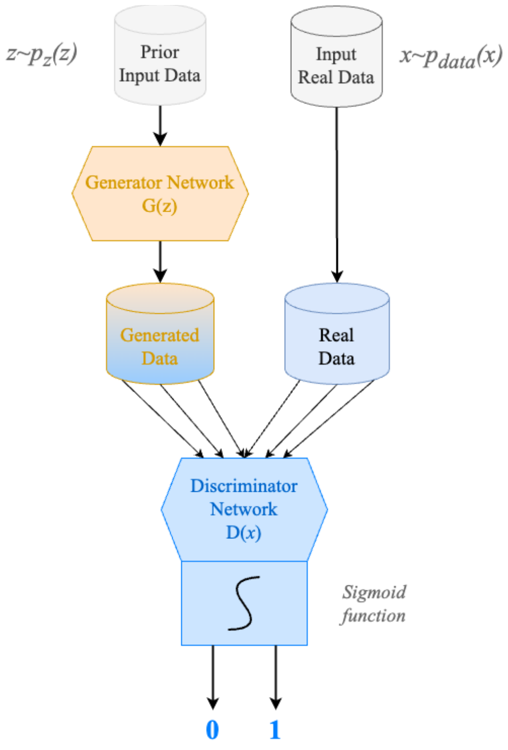
While SMOTE generates synthetic samples efficiently for traditional ML models, GANs produce realistic, high-quality data for complex, imbalanced datasets. Together with Grid Search Cross Validation (CV), these techniques enhance the model's performance, achieving superior accuracy and recall (Chawla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

F. Model Design and Implementation (Figure 2)

1) Step 1: Data Preparation and Processing

Features such as age, cholesterol, and blood pressure are preprocessed to ensure data integrity. Missing values are handled using imputation methods, and features are normalized with the Standard Scaler for consistency, a critical step for neural networks (Pedregosa et al., 2011).

Figure 1- Architecture of the GAN network



2) Step 2: GAN Architecture Definition

GANs generate synthetic heart failure data through two networks: (a) Generator Network: Takes a latent vector (random noise) and produces synthetic patient data via fully connected layers with ReLU activations. Tanh activation ensures compatibility with medical data. (b) Discriminator Network: Classifies real and synthetic data using LeakyReLU and Sigmoid activations, ensuring the synthetic data closely resembles real patient profiles (Goodfellow et al., 2014).

3) Step 3: GAN Training

The GAN is trained over 5,000 epochs using the Adam optimizer (learning rate = 0.00005). The Generator creates realistic data while the Discriminator improves its classification ability, resulting in high-quality synthetic data.

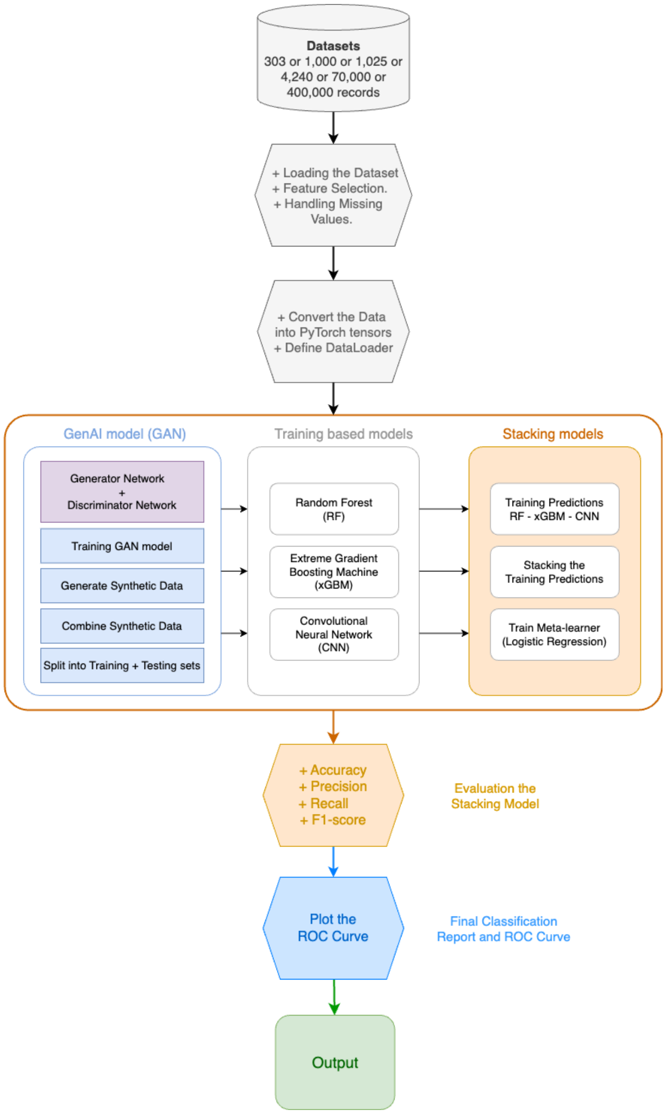
4) Step 4: Synthetic Data Generation

The trained Generator produces synthetic samples, which are merged with the original dataset to create a larger, diverse training set. This improves model generalization by introducing variability.

5) Step 5: Data Splitting and Normalization

The expanded dataset (real + synthetic data) is split into training (80%) and test (20%) sets. Standard scaling ensures normalized features, optimizing neural network learning.

Figure 2- Architecture of the Stacking Generative AI model


Datasets
303 or 1,000 or 1,025 or 4,240 or 70,000 or 400,000 records
+ Loading the Dataset
+ Feature Selection.
+ Handling Missing Values.
+ Convert the Data into PyTorch tensors
+ Define DataLoader
GenAI model (GAN)
Generator Network + Discriminator Network
Training GAN model
Generate Synthetic Data
Combine Synthetic Data
Split into Training + Testing sets
Training based models
Random Forest (RF)
Extreme Gradient Boosting Machine (xGBM)
Convolutional Neural Network (CNN)
Stacking models
Training Predictions RF - xGBM - CNN
Stacking the Training Predictions
Train Meta-learner (Logistic Regression)
+ Accuracy
+ Precision
+ Recall
+ F1-score
Evaluation the Stacking Model
Plot the ROC Curve
Final Classification Report and ROC Curve
Output

6) *Step 6: Base Model Training*

Base models include:

Random Forest (RF): Configured with 100 trees, a maximum depth of 10, and a minimum split size of 10.

Extreme Gradient Boosting (xGBM): Configured with 200 estimators, a 0.05 learning rate, and a subsample ratio of 0.8 for feature interaction modeling.

Convolutional Neural Network (CNN): Incorporates a Conv1D layer (16 filters, kernel size = 3), MaxPooling, and Dropout layers to prevent overfitting. It uses binary cross-entropy loss, the Adam optimizer, and early stopping.

7) *Step 7: Meta-Learner Training*

Predictions from RF, xGBM, and CNN serve as inputs for the stacking model's meta-learner, implemented with Logistic Regression. The meta-learner combines base model outputs to make final predictions.

8) *Step 8: Model Evaluation*

The stacked model is evaluated using accuracy, precision, recall, F1-score, and ROC AUC. The ROC curve highlights sensitivity and specificity balance, with a high AUC indicating excellent predictive performance.

9) *Step 9: Final Report*

A classification report details metrics for each class, while the ROC curve demonstrates the model's robustness and clinical applicability (Figure 2).

The Stacking Generative AI Model integrates traditional ML techniques with DL models and GAN-generated synthetic data, achieving exceptional predictive accuracy. Steps 1 to 9 ensure the model is robust, scalable, and clinically reliable, making it ideal for early heart failure detection and management. This approach represents a significant advancement in predictive analytics for healthcare (Chawla et al., 2002; Goodfellow et al., 2014; Pedregosa et al., 2011; Radford et al., 2015).

G. *Evaluation Measurement and Validation Methods*

The performance of the Stacking Generative AI model was evaluated using a combination of metrics, including accuracy, ROC AUC, precision, recall, and F1-score. These metrics ensure a comprehensive understanding of the model’s classification capabilities.

1) *Cross-Validation*

K-fold cross-validation, including 5- and 10-fold methods, was used to validate the model across multiple data splits. This resampling technique ensures the model’s robustness by training on k−1 folds and testing on the remaining fold iteratively. For 5-fold cross-validation, the model achieved accuracies of [0.9938, 1.0000, 0.9877, 0.9969, 0.9938], resulting in a mean accuracy of 99.4%. Similarly, 10-fold cross-validation confirmed the model's consistency with the same mean accuracy. These results demonstrate the reliability and generalization of the model across unseen data (James et al., 2013).

2) *Hyperparameter Tuning*

Grid search optimized the hyperparameters of the Logistic Regression meta-learner by systematically exploring values for C, the inverse of regularization strength. The best C=0.01 was selected based on cross-validation results, enhancing the model’s predictive performance.

3) *Comprehensive Validation*

The combination of cross-validation, learning curves, regularization, and hyperparameter tuning ensured that the Stacking Generative AI model is robust and well-calibrated. These techniques minimized overfitting while maximizing generalization, making the model suitable for deployment in heart failure prediction scenarios.

IV. Results

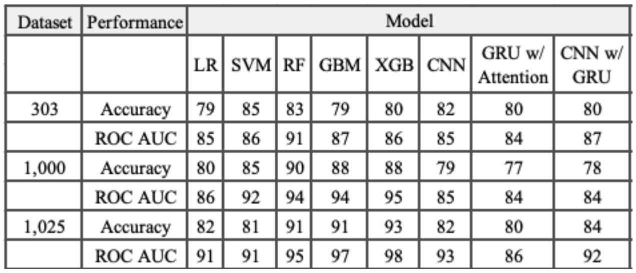
A. *Performance Comparison between Traditional Models and Neural Network Models: How do traditional ML models compare to neural network-based models in terms of accuracy and ROC AUC for heart failure prediction?*

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

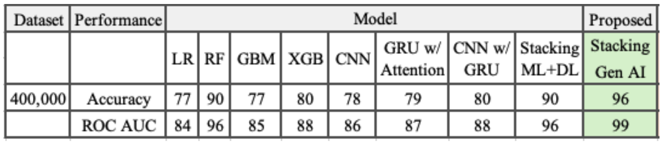


Dataset	Performance	Model							
		LR	SVM	RF	GBM	XGB	CNN	GRU w/ Attention	CNN w/ GRU
303	Accuracy	79	85	83	79	80	82	80	80
	ROC AUC	85	86	91	87	86	85	84	87
1,000	Accuracy	80	85	90	88	88	79	77	78
	ROC AUC	86	92	94	94	95	85	84	84
1,025	Accuracy	82	81	91	91	93	82	80	84
	ROC AUC	91	91	95	97	98	93	86	92

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

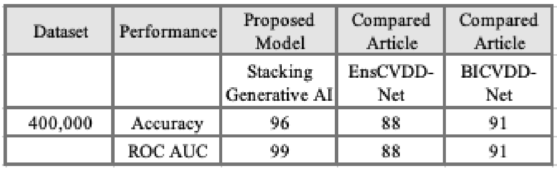


Dataset	Performance	Model								Proposed
		LR	RF	GBM	XGB	CNN	GRU w/ Attention	CNN w/ GRU	Stacking ML+DL	Stacking Gen AI
400,000	Accuracy	77	90	77	80	78	79	80	90	96
	ROC AUC	84	96	85	88	86	87	88	96	99

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

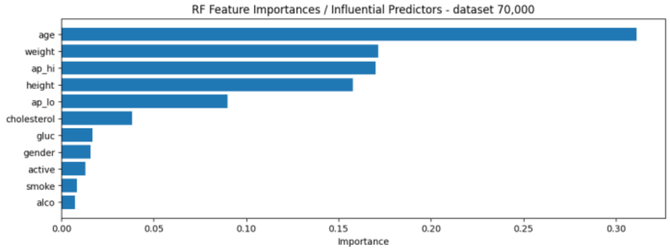
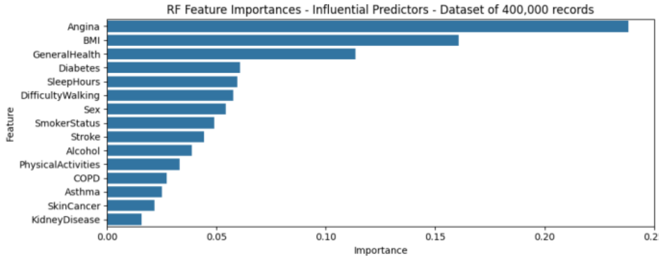


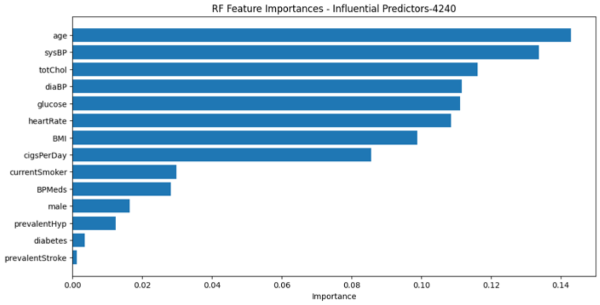
Figure 4 –Influential Predictors – dataset of 400,000 records



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. On another hand, the 11,627-record dataset, HDL cholesterol, age, and sysBP emerged as significant features, leading to 91% accuracy and a 0.95 ROC AUC.

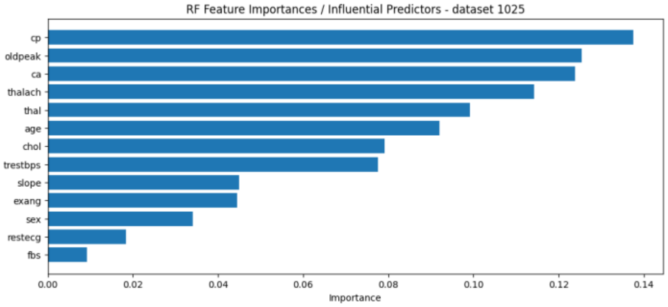
Figure 5 – Influential Predictors – dataset of 4,240 records



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 a novel 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.

Table 4 – Results across 12 models with evaluation on 7 datasets range from 303 – 400,000 records.

Dataset	Performance	Model										Proposed Model	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
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	
1,000	Accuracy	80	85	90	88	88	79	77	78	94	98	98	94	Dumlao, J. (n.d.)
	ROC AUC	86	92	94	94	95	85	84	84	98	99	99.9	NA	
1,025	Accuracy	82	81	91	91	93	82	80	84	95	95	98	85	Nasser, A. (n.d.)
	ROC AUC	91	91	95	97	98	93	86	92	98	99	99.9	NA	
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	
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	
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	
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	

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.

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, and 9).

V. Conclusion

This study demonstrates the efficacy of the Stacking Generative AI model as a hybrid solution for heart disease prediction. By integrating Generative AI with traditional ML models like RF, GBM, and xGBM, alongside DL architectures such as CNNs and RNNs, the model addresses key challenges in predictive modeling. These include handling class imbalance, improving scalability, and achieving superior predictive accuracy across datasets of varying sizes and complexities.

A. Summary of Findings

The Stacking Generative AI model consistently outperformed standalone ML and DL models across datasets ranging from 303 to 400,000 records. On the 1,000-record dataset, it achieved an ROC AUC of 0.999, surpassing standalone xGBM (0.94) and CNN (0.85). Even on the largest dataset of 400,000 records, the model maintained high performance with an ROC AUC of 0.99 and 96% accuracy (Figure 9), confirming its scalability and robustness for large-scale clinical applications.

The model's hybrid structure combines the structured data analysis capabilities of ML models with the complex pattern recognition abilities of DL models. GANs play a critical role in balancing minority classes by generating synthetic data, improving recall and F1-scores. This feature is particularly valuable in healthcare, where imbalanced datasets are common, and accurate prediction of high-risk cases is essential.

Figure 7- ROC AUC performance on dataset of 303 records

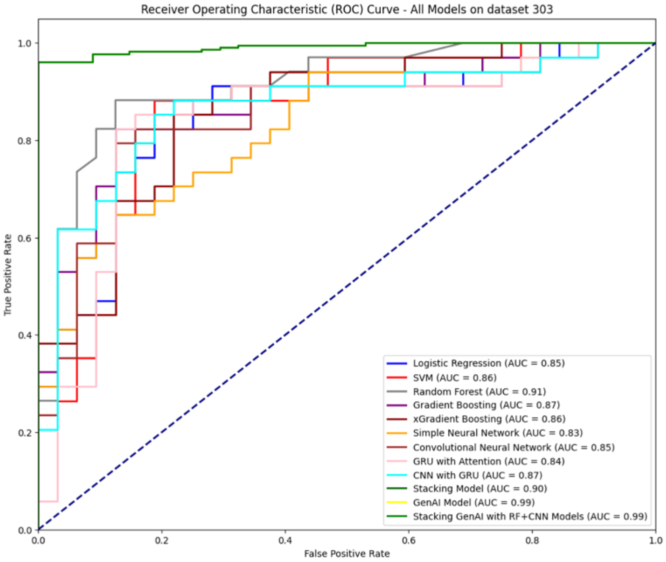


Figure 8- ROC AUC performance on dataset of 11,627 records

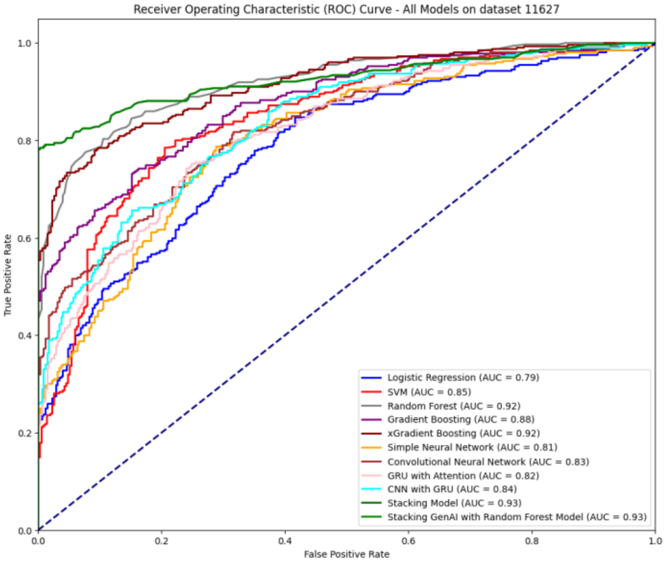
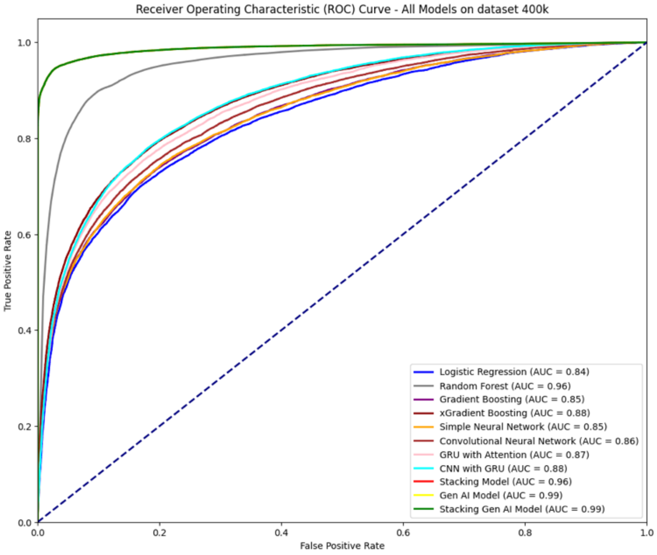


Figure 9- ROC AUC performance on dataset of 400,000 records



B. *Comparison with Existing Literature*

Compared to existing studies, the proposed model demonstrates clear advantages. Singh et al. (2024) reported an ROC AUC of 0.89 using xGBM, while the Stacking Generative AI model achieved 0.99 on comparable datasets. Similarly, the model outperformed RF-based approaches from Chicco et al. (2022), which achieved an ROC AUC of 0.85. These results underscore the effectiveness of integrating GANs with ML and DL techniques in a stacking framework.

C. *Clinical Implications*

The Stacking Generative AI model has significant potential in clinical applications. Its ability to handle imbalanced datasets and deliver high predictive accuracy makes it a reliable tool for early diagnosis and personalized treatment planning. Key predictors identified by the model, such as systolic blood pressure, BMI, total cholesterol, and glucose levels, provide actionable insights for clinicians, enabling better resource allocation and patient care.

The model's adaptability to diverse datasets further enhances its utility, ensuring generalizability across different clinical settings. The largest improvement was observed model's consistent performance on datasets of various sizes, from 303 to 400,000 records, highlights its robustness and scalability. While complex models like the Stacking Generative AI model may lack the interpretability of simpler approaches like LR, the trade-off is justified by its superior predictive power, which can serve as a decision-support tool for clinicians.

D. *Limitations and Future Research*

The model's application to datasets from diverse geographical or clinical settings remains a limitation. While it performed well across the datasets used in this study, additional research is needed to validate its generalizability to broader populations. Furthermore, the complexity of the model may pose challenges for direct integration into clinical workflows, necessitating user-friendly interfaces for clinicians.

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. Moreover, 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. This research paves the way for future advancements in healthcare predictive modeling, bridging the gap between academic innovation and clinical practice.

VI. Discussion and Future Works

A. *Discussion*

The Stacking Generative AI model advances HF prediction by integrating ML, DL, and GAN techniques. It effectively handles imbalanced datasets, enhances scalability, and improves accuracy, achieving results such as 95% accuracy and a 0.99 ROC AUC on small datasets and maintaining 96% accuracy and a 0.99 ROC AUC on large datasets. Compared to standalone RF, CNN, and xGBM models, it demonstrates significant improvements, offering robust and generalizable predictions across diverse datasets.

While its results align with prior studies, the hybrid approach extends performance by leveraging SHAP and LIME for interpretability and improving usability for clinical decision-making. Clinically, the model facilitates early HF detection, personalized care, and resource optimization. However, future work is needed to address computational demands and validate the model in real-world healthcare settings.

B. Future Works

Exploring Advanced Models: Incorporate transformer-based models and reinforcement learning for enhanced performance in sequential tasks. Integrating large language models with EHRs may further improve predictions.

Application Development: Create web and mobile platforms for real-time HF monitoring to improve accessibility for clinicians and patients.

Expanding Applications: Adapt the model for other medical conditions like diabetes and chronic kidney disease to demonstrate broader utility.

Enhancing Interpretability: Develop counterfactual explanations and visualization tools for better clinical usability.

Real-World Validation: Conduct clinical trials to refine the model's design and usability based on feedback from healthcare practitioners.

Improving Efficiency: Apply techniques like model pruning, quantization, and edge computing to optimize performance in resource-limited environments.

Incorporating Diverse Data: Future research should include genomic, imaging, and patient-reported data to create multimodal models, providing a holistic view of patient health and identifying novel biomarkers (Chen et al., 2023).

C. Summary

The Stacking Generative AI model represents a major step forward in predictive healthcare. Its integration of ML, DL, and GANs addresses key challenges such as class imbalance, scalability, and accuracy, making it a powerful tool for HF prediction and management. Future efforts should focus on expanding its applications, improving its efficiency, and validating its real-world utility to maximize its 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] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

[7] Pedregosa et al. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research, 12 (2011): 2825-2830.

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

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

[10] James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning: with Applications in R. Springer, 2013.

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

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

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

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

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

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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).

[25] Chen, H., et al. (2023). Hyperparameter tuning in healthcare models. International Journal of Data Science, 19(1), 90-110.