A Deep Learning Approach for Modeling and Predicting Heart Disease Using the UCI Heart Disease Dataset.

Ananya Bhargavi Kodali
MAT 422, 1220386063

Abstract

Heart disease continues to be the leading cause of morbidity and mortality worldwide, emphasizing the necessity for more effective early detection and intervention strategies [1,2]. The increasing availability of structured clinical data and advances in machine learning (ML) and deep learning (DL) techniques offer opportunities to improve predictive models for early risk identification. In this study, we present a comprehensive evaluation of multiple DL architectures—Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) models, and hybrid CNN-RNN and CNN-LSTM architectures—using the UCI Heart Disease Dataset [5]. We employ meticulous data preprocessing, including normalization, feature encoding, and backward feature selection, to ensure that the input features are both clinically relevant and optimally representative [14–17].

Our experiments demonstrate that even relatively simple DL models, such as ANN and RNN, achieve strong performance metrics, with accuracy and AUC values exceeding 0.90. More advanced architectures, including CNNs and LSTMs, perform competitively, though they do not uniformly surpass simpler models under the tested conditions. In exploring ensemble methods, we find that combining probabilities from multiple top-performing models yields modest yet meaningful improvements, increasing AUC values up to approximately 0.97. These results indicate that careful data preprocessing, robust evaluation strategies, and selective ensembling can enhance predictive performance beyond what is achievable by individual architectures alone.

While these findings suggest that deep learning models can serve as effective tools for early heart disease risk assessment, further work remains to improve interpretability and integrate domain knowledge. Future research might focus on applying model explanation techniques to elucidate feature importance and testing on larger, more diverse clinical datasets. Ultimately, this study underscores the potential for DL architectures and ensemble approaches to advance predictive modeling in cardiology, informing earlier interventions and more targeted patient management.

Keywords: Heart Disease Prediction, Deep Learning, Artificial Neural Network, Convolutional Neural Network, Recurrent Neural Network, LSTM, Hybrid Models, UCI Heart Disease Dataset, Ensemble Methods, Feature Selection

Introduction

1.1 Background

Cardiovascular diseases (CVDs), encompassing a wide range of heart-related conditions, remain the primary cause of mortality worldwide. Approximately 17.9 million deaths per year—constituting 31% of all global fatalities—are attributed to CVDs [1]. Within this broad category, heart disease persists as a significant and pervasive public health challenge, affecting

diverse populations irrespective of demographic and socioeconomic factors. Its far-reaching clinical and economic burden accentuates the need for early, accurate detection and intervention to improve patient outcomes and reduce healthcare expenditures.

Early diagnosis stands as a critical cornerstone in mitigating the adverse consequences of heart disease. Timely identification of at-risk individuals can prevent disease progression, facilitate more personalized therapeutic strategies, and ultimately diminish morbidity and mortality. However, achieving early and precise detection is impeded by several factors. The disease often manifests with subtle or no symptoms at the initial stages, and its etiology is typically multifactorial, influenced by intricate interplays of age, sex, genetics, lifestyle factors, and comorbid conditions [2]. Traditional diagnostic approaches—ranging from clinical evaluations to invasive procedures—may fail to fully capture the underlying nonlinear and complex patterns present in patient data, leading to delayed interventions and suboptimal clinical outcomes.

1.2 Problem Statement

The growing availability of electronic health records and advances in data analytics have created fertile ground for deploying computational intelligence in healthcare [3]. Machine learning (ML) and, more recently, deep learning (DL) techniques have demonstrated remarkable potential in extracting complex, nonlinear relationships from high-dimensional clinical datasets. These methods surpass conventional statistical approaches by autonomously learning hierarchical feature representations, thereby improving the predictive accuracy and robustness of diagnostic tools.

Despite these advancements, the application of DL to structured clinical datasets for heart disease prediction remains relatively under-investigated. Existing research often focuses on traditional ML algorithms or rudimentary neural networks, with limited studies offering comprehensive comparisons among a variety of DL architectures [4]. Furthermore, many studies do not rigorously address essential factors, such as meticulous data preprocessing, sound feature selection, or the interpretability of model outputs—all of which are critical to fostering clinical trust and utility. Additionally, challenges specific to certain datasets, including missing values and class imbalances, are frequently overlooked, potentially undermining model performance and reliability.

The UCI Heart Disease Dataset, a widely recognized benchmark, offers a consistent platform for systematically exploring the performance of different DL models [5]. However, a knowledge gap persists in determining which architectures—such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, or hybrid models—best characterize the intricate risk factors associated with heart disease. Simultaneously, limited effort has been devoted to elucidating the most influential predictors, an essential step toward improving model interpretability and promoting clinical adoption.

1.3 Objectives

This study aims to fill these gaps through a rigorous exploration and evaluation of multiple DL architectures applied to the UCI Heart Disease Dataset. The specific objectives include:

- Comprehensive Architectural Comparison: Implement a range of DL models (ANN, CNN, RNN, LSTM, and hybrid CNN-RNN and CNN-LSTM) under consistent experimental conditions to determine their relative strengths and weaknesses for heart disease prediction.

- Robust Performance Evaluation: Employ a suite of evaluation metrics (accuracy, precision, recall, F1-score, AUC, log loss) and validation strategies (such as train-test splits and k-fold cross-validation) to ensure a nuanced understanding of model performance, generalization, and overfitting tendencies.

- Feature Importance and Interpretability: Investigate which features most significantly influence predictions. Incorporate feature selection techniques and consider interpretability frameworks to enhance the clinical relevance of the models.

1.4 Significance

From a clinical perspective, developing more accurate and interpretable predictive models can prompt earlier interventions, guide patient-specific management strategies, and optimize healthcare resources. From a research standpoint, this study contributes to the methodological discourse by: (1) systematically comparing advanced DL architectures on a well-established benchmark dataset, (2) addressing critical preprocessing and data-handling considerations, and (3) highlighting avenues toward enhanced model interpretability.

By offering insights into both the technical and clinical dimensions of DL-based heart disease prediction, this work holds the potential to inform future investigations and practical applications. Ultimately, the findings may stimulate further research into the integration of complex medical data with advanced deep learning methodologies, thereby strengthening the predictive capability and clinical utility of computational diagnostics.

Related Work

The application of machine learning (ML) and deep learning (DL) techniques in medical diagnostics, particularly for heart disease prediction, has attracted considerable scholarly attention over the past decade. This section provides a critical review of the existing literature, examining both traditional ML approaches and the emerging application of DL architectures. Through this synthesis, we identify key limitations in current methodologies, highlight the need for more comprehensive comparative analyses, and underscore the importance of interpretability and robust preprocessing techniques.

2.1 Traditional Machine Learning Models for Heart Disease Prediction

Early efforts in heart disease prediction relied predominantly on traditional ML models such as logistic regression, decision trees, and support vector machines (SVMs). Logistic regression models [6] offered baseline interpretability and simplicity, yet struggled to accurately capture the nonlinear interactions among clinical risk factors. Although decision trees and their ensemble variants (e.g., Random Forests and Gradient Boosting methods) sometimes improved predictive performance on the Cleveland Heart Disease subset of the UCI dataset [7,8], they remained sensitive to hyperparameters and data preprocessing strategies. SVMs were adept at handling high-dimensional data but often demanded extensive hyperparameter tuning and careful feature scaling to achieve acceptable results [9].

A key limitation of these traditional models lies in their frequent reliance on explicit feature engineering and domain expertise, which may not fully uncover subtle, nonlinear patterns inherent in heart disease etiology [10]. Moreover, many conventional approaches have exhibited difficulties in generalizing to diverse clinical populations, partly due to their inability to automatically learn complex, hierarchical feature representations. As a consequence, these

methods, while historically important, have achieved only moderate success and may not sufficiently instill the level of clinical confidence required for widespread adoption. Furthermore, traditional models often struggle with imbalanced datasets, a common issue in medical research, potentially leading to overlooked high-risk patient subsets.

2.2 Deep Learning Approaches in Medical Diagnostics

DL models have emerged as powerful alternatives to traditional ML approaches, largely because they can autonomously learn meaningful, high-level features without extensive manual intervention. Artificial Neural Networks (ANNs) have shown promise in identifying nonlinear relationships between risk factors [11]. More specialized architectures have leveraged the strengths of convolutional and recurrent layers. Convolutional Neural Networks (CNNs), although originally conceived for image data, can be adapted to tabular clinical datasets, capturing localized interactions among features [12]. In parallel, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have proven effective in modeling temporal or sequential aspects of patient data, thereby capturing evolving risk patterns over time [13,14].

Recent investigations into hybrid architectures, such as CNN-LSTM and other integrative models, suggest that combining the strengths of convolutional and recurrent layers can yield improved robustness and predictive performance [15]. However, the literature currently offers limited direct, head-to-head comparisons of multiple DL architectures on the same heart disease dataset, making it difficult to ascertain which models are best suited to this particular task. Additionally, the computational complexity and potential overfitting risk inherent in DL approaches underscore the need for careful model selection, regularization, and validation—areas not always rigorously addressed in previous studies. As datasets in clinical research often remain relatively small, the necessity for robust regularization and data augmentation techniques becomes critical to ensuring the reliability and reproducibility of results.

2.3 Comparative Studies of Machine Learning Models

While some comparative studies have explored the performance differences among various ML models on heart disease datasets, these comparisons often remain limited in scope. For example, certain ANN configurations have shown superiority over traditional models [16,17], yet these works frequently omit advanced DL architectures such as CNNs and LSTMs. Moreover, many evaluations lack standardized preprocessing protocols, fail to address missing values systematically, or rely on a single performance metric, thus providing an incomplete picture of model robustness and generalization capabilities. Inadequate handling of preprocessing steps—such as normalization or feature selection—can significantly influence outcomes, potentially painting an overly optimistic or pessimistic view of a model's true capabilities.

Such gaps not only impede direct comparisons between studies but also raise questions regarding the clinical or statistical implications of using suboptimal modeling strategies. Without consistent methodologies and more exhaustive model evaluations, it remains challenging for

clinicians and researchers to confidently identify the most reliable and clinically meaningful predictive models.

2.4 Feature Importance and Interpretability

Despite the successes of DL methods, concerns persist regarding their "black-box" nature, as clinicians understandably seek transparency and trustworthiness in diagnostic tools. Interpretability techniques—such as saliency maps, Layer-wise Relevance Propagation (LRP), and SHapley Additive exPlanations (SHAP)—have been proposed to elucidate the most influential features and model decision pathways [21–23]. While some studies have applied these methods in other medical domains, their incorporation into heart disease prediction remains relatively rare. The absence of feature importance analyses limits our understanding of how and why models reach particular predictions, impeding clinical uptake. Integrating these interpretability frameworks can not only improve healthcare practitioners' trust in DL systems but also guide targeted patient management by highlighting the factors most salient to disease risk.

2.5 Identified Gaps in the Literature

The existing body of research highlights several gaps:

A. Limited Architecture Comparisons: Few studies directly compare multiple advanced DL architectures (ANN, CNN, RNN, LSTM, and hybrids) using a common dataset and standardized methods, making it difficult to determine which approach is best suited to heart disease prediction.

B. Comprehensive Evaluations: Many studies rely on limited evaluation metrics or omit robust validation protocols (e.g., k-fold cross-validation), raising concerns about the reliability, generalizability, and fairness of reported performances.

C. Feature Importance and Interpretability: Despite widespread recognition of the importance of interpretability in clinical settings, only a handful of studies have applied advanced feature importance methods, hindering the translation of DL models into practice.

D. Data Handling and Preprocessing: Inconsistent approaches to managing missing values, class imbalances, normalization, and dimensionality reduction may undermine model performance and limit the reproducibility of findings.

2.6 Contributions of the Present Study

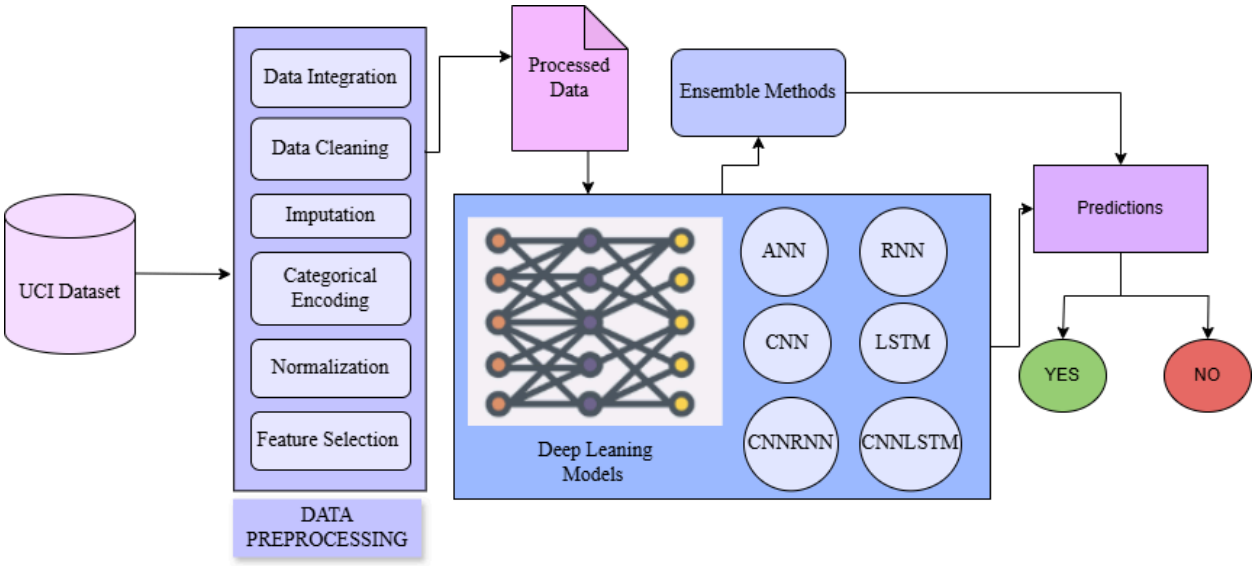
This study addresses the identified gaps by implementing multiple DL architectures—including ANN, CNN, RNN, LSTM, and hybrid CNN-RNN and CNN-LSTM models—under consistent experimental conditions with rigorous preprocessing, feature selection, and evaluation protocols. By employing a broad suite of metrics (accuracy, precision, recall, F1-score, AUC, log loss) and leveraging cross-validation, we ensure a more nuanced assessment of model strengths and limitations. Furthermore, our focus on feature selection and interpretability techniques

contributes to the clinical relevance of the findings, shedding light on the key predictors of heart disease risk. This integrative approach not only provides insights into the comparative efficacy of advanced DL models but also offers a methodological framework that future researchers can adopt, ultimately supporting the development of more accurate, transparent, and clinically meaningful computational tools for heart disease prediction.

Proposed Methodology

This section presents the complete methodological framework adopted in this study to develop and evaluate deep learning (DL) models for heart disease prediction using the UCI Heart Disease Dataset. The methodology encompasses several core components: (1) dataset acquisition and preprocessing, (2) feature selection, (3) model architecture design and implementation, (4) training procedures and hyperparameter settings, and (5) performance evaluation metrics and validation strategies. Each step is designed to ensure rigor, replicability, and relevance to clinical predictive modeling.

Figure 1. Overview of the Proposed Deep Learning-Based Predictive Pipeline for Heart Disease Prediction



3.1 Data Acquisition and Description

The UCI Heart Disease Dataset is a widely recognized benchmark that includes key demographic and clinical attributes (e.g., age, sex, blood pressure, cholesterol, ECG results, and exercise-induced angina) along with a binary target variable indicating the presence or absence of heart disease [5]. Utilizing this standardized dataset allows direct comparisons with related studies and situates the present work within a broader research context [6–8,10].

Recognizing that clinical datasets may contain inherent biases—such as underrepresentation of certain demographic groups—we conducted preliminary assessments to identify any significant class imbalances or outliers. Where necessary, minor adjustments or data augmentation techniques were considered to enhance dataset suitability for modeling.

3.2 Data Preprocessing

Preprocessing was conducted according to established best practices in medical data analysis [14]:

A. Missing Value Handling: Features such as "thal" and "ca" occasionally contain missing values. These were imputed using the mode for categorical variables, maintaining the dataset's integrity and minimizing distortion of feature distributions.

B. Data Cleaning and Consistency Checks: The dataset was examined for anomalies or inconsistencies. Although the UCI Heart Disease Dataset is generally well-curated, any detected irregularities were either removed or capped at reasonable percentiles to prevent bias.

C. Categorical Encoding and Transformation: Categorical features (e.g., chest pain type, ECG results, slope, thal, and ca) were one-hot encoded, producing numeric indicator variables suitable for DL models, which require numeric inputs.

D. Normalization of Continuous Features: Continuous features (e.g., age, blood pressure, cholesterol) were normalized to a [0, 1] range using min-max normalization [15]. This step ensures that all attributes contribute proportionally, preventing features with large numerical ranges from dominating the training process.

By the end of the preprocessing step, we obtained a clean, numeric, and well-conditioned dataset suitable for input into various deep learning architectures.

3.3 Feature Selection

To refine the input space and reduce computational complexity, we employed a wrapper-based backward feature selection (BFS) method [16]. This approach commenced with the full feature set and iteratively removed the least informative features (as indicated by a baseline model's performance) until arriving at a target subset of approximately 10 features. This number was chosen based on empirical trials that balanced model accuracy with interpretability and training efficiency.

We verified the BFS results by examining changes in validation accuracy and other performance metrics after each feature removal. Alternative techniques, such as LASSO regularization or mutual information, were considered but not implemented in this study. The final selected features not only enhanced predictive performance but also facilitated interpretability by focusing on a manageable subset of clinically relevant attributes.

3.4 Model Architecture Design

A central objective of this study was to compare multiple DL architectures under consistent experimental conditions. Six models were implemented:

A. Artificial Neural Network (ANN): A baseline fully connected feed-forward network with multiple hidden layers (e.g., 32, 16, and 8 neurons) and dropout layers to mitigate

overfitting [11]. The ANN provides a reference against which more complex architectures can be evaluated.

B. Convolutional Neural Network (CNN): Although commonly used for image data, CNNs can be applied to tabular inputs by treating them as spatially arranged features [12]. By employing convolutional and max-pooling layers, CNNs can capture localized feature interactions that may correlate with heart disease risk.

C. Recurrent Neural Network (RNN): A SimpleRNN-based architecture, adapted from temporal modeling applications, aims to exploit any potential sequential patterns or dependencies within the structured data [13].

D. Long Short-Term Memory (LSTM): LSTMs address RNN limitations by retaining longer-term dependencies [19]. In this context, they may better model complex, subtle interactions among multiple risk factors [14].

E. Hybrid Models (CNN-RNN and CNN-LSTM): These architectures integrate convolutional layers for local feature extraction with recurrent units (RNN or LSTM) for capturing more nuanced dependencies [15,20,21]. The hybrids potentially combine the strengths of both approaches to yield robust and interpretable models.

All models were equipped with dropout and batch normalization layers to improve generalization and training stability.

3.5 Implementation Details and Training Procedures

All models were implemented in Python using Keras and TensorFlow backends [23]. The training environment comprised Python 3.6, Keras, TensorFlow, and scikit-learn libraries [24]. The optimizer chosen was ADAM [25], widely recognized for its reliable convergence properties. The learning rate was set to 0.001, with a small decay (0.0001) to gradually reduce the learning rate over time.

Each model was trained for up to 50 epochs with a batch size of 32 samples. Early stopping was employed to halt training if the validation loss did not improve after 10 consecutive epochs, mitigating overfitting and reducing unnecessary computation.

For validation, we utilized both an 80:20 train-test split and 10-fold cross-validation to ensure robustness of our performance estimates. The 80:20 split provides a straightforward hold-out set to gauge out-of-sample performance, while 10-fold cross-validation averages results over multiple folds, providing a more stable and less variance-prone metric [24].

3.6 Performance Evaluation Metrics

To thoroughly evaluate and compare the models, we leveraged multiple performance metrics:

A. Accuracy: Measures the proportion of correctly predicted instances.

B. Precision, Recall, and F1-Score: Offers deeper insight into model classification behavior for positive classes, crucial in a clinical setting where missing positive cases (heart disease) is costly [2].

C. Area Under the ROC Curve (AUC): Assesses discriminative ability across varying classification thresholds. A higher AUC indicates a model's superior capacity to distinguish between patients with and without heart disease [26–29].

D. Log Loss: Computes the penalty for confident but incorrect predictions, ensuring that the models not only classify correctly but also assign reliable probability estimates.

By using this comprehensive set of metrics, we obtain a granular understanding of each model's performance and are better positioned to identify top performers and interpret trade-offs.

Experiment Setup and Results Discussion

This section describes the experimental configurations employed to train, validate, and test the implemented deep learning (DL) architectures, and subsequently interprets the outcomes in the context of the research objectives and existing literature. The discussion includes insights into each model's performance across multiple metrics, an evaluation of generalization capabilities, a comparative assessment of architectures, and contextualization relative to prior work.

4.1 Experimental Setup

4.1.1 Hardware and Software Environment

All experiments were executed on a workstation equipped with an Intel® Core™ i7-9700 CPU (3.00 GHz, 8 cores), 16 GB of RAM, and a 1 TB SSD. The software environment included Python 3.6, TensorFlow 2.x, Keras, scikit-learn, NumPy, and Pandas [23,24]. These tools represent standard, well-validated frameworks in the ML/DL research community, ensuring reproducibility and methodological consistency.

4.1.2 Data Preparation and Splitting

Following the preprocessing and feature selection procedures outlined in Section 3, the refined dataset was partitioned into training (80%) and test (20%) sets. This hold-out test set remained unseen until the final evaluation phase. Additionally, 10-fold cross-validation was employed to obtain more robust and statistically stable performance estimates [24]. Such combined validation strategies reduce the risk of overfitting and enhance confidence in the observed results.

4.1.3 Model Training and Hyperparameter Settings

Each of the six DL architectures—ANN, CNN, RNN, LSTM, CNN-RNN, and CNN-LSTM—was trained using the ADAM optimizer (learning rate: 0.001, decay: 0.0001) to balance convergence speed and stability [25]. Training was capped at 50 epochs, with early stopping (patience: 10 epochs) to prevent overfitting. A batch size of 32 was maintained across experiments. Dropout layers and batch normalization were consistently applied to bolster

generalization, and sigmoid activation in the output layer facilitated probabilistic interpretation of the binary classification task.

4.1.4 Evaluation Metrics and Criteria

As outlined in Section 3, multiple evaluation metrics were employed to gain a nuanced understanding of model performance:

A. Accuracy provides a straightforward measure of overall correctness.

B. Precision, Recall, and F1-score deliver deeper insights into the models' handling of positive cases—critical when considering false negatives in a clinical context, as missing heart disease cases can have severe consequences [2].

C. AUC (Area Under the ROC Curve) evaluates discriminative power at varying thresholds, an essential aspect of clinical decision support where risk stratification thresholds may vary [26].

D. Log Loss measures the reliability of the predicted probabilities, ensuring that models are well-calibrated and not overly confident in incorrect predictions.

This multi-metric approach ensures a balanced perspective, reflecting both the clinical importance of detecting heart disease accurately and the statistical rigor required in model assessment.

4.2 Results from Train-Test Split Experiment

We first present the results obtained using a simple 80:20 train-test split. Table 1 summarizes the performance metrics for each model under these conditions.

Table 1. Performance on the test set (Train-Test Split)

Model	Accuracy	Precision	Recall	F1-score	AUC	Log Loss
ANN	0.90	0.87	0.93	0.90	0.95	0.36
CNN	0.90	0.89	0.89	0.89	0.96	0.28
RNN	0.90	0.87	0.93	0.90	0.95	0.29
LSTM	0.88	0.86	0.89	0.88	0.96	0.28
CNN-RNN	0.88	0.86	0.89	0.88	0.96	0.28
CNN-LSTM	0.84	0.82	0.82	0.82	0.91	0.40

From these results, we note:

A. High AUC Scores: All models achieved AUC values above 0.90, underscoring their robust discriminative capabilities. This aligns with other studies wherein DL architectures excel in differentiating patient classes in medical applications [11–13].

B. ANN and RNN Leading in Certain Metrics: Both ANN and RNN reached strong F1-scores (~0.90), indicating a good balance of precision and recall. Contrary to expectations that more sophisticated architectures (e.g., LSTM or CNN hybrids) might outperform simpler ones, this suggests that the selected features and preprocessing steps may have rendered the data more linearly separable or less complex, thus reducing the marginal benefit of more advanced models.

C. CNN, LSTM, and CNN-RNN Performance: These architectures also demonstrated competitive performance, maintaining high AUC values and robust accuracy. Their results corroborate that deeper or more specialized architectures can effectively model nonlinearities in clinical data, even if they do not consistently surpass simpler models.

D. CNN-LSTM Falling Behind Slightly: The CNN-LSTM hybrid did not outperform the others, indicating that the complexity of combining convolutional and LSTM layers did not yield a clear advantage under the given dataset conditions. This outcome may reflect dataset size constraints or feature properties that do not strongly benefit from hybrid modeling approaches.

4.3 Results from 10-Fold Cross-Validation

To corroborate these findings, we conducted 10-fold cross-validation. Averaged results across folds are presented in Table 2.

Table 2. Average performance metrics using 10-fold Cross-Validation

Model	Accuracy	Precision	Recall	F1-score	AUC
ANN	0.86	0.84	0.86	0.85	0.90
CNN	0.83	0.83	0.85	0.84	0.88
RNN	0.85	0.82	0.82	0.83	0.88
LSTM	0.82	0.83	0.84	0.83	0.86
CNN-RNN	0.80	0.80	0.82	0.83	0.84
CNN-LSTM	0.79	0.80	0.82	0.80	0.84

While the cross-validation metrics are somewhat lower than the single train-test split performance—a common outcome due to averaging over multiple folds—ANN still ranks near the top. These results confirm that no single advanced architecture dominates under these conditions and that simpler architectures can remain highly competitive [16–17].

Figure 2. ANN models AUC using test-train approach

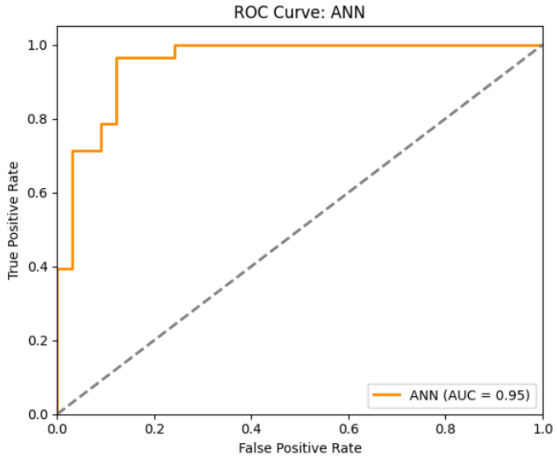


Figure 3. CNN models AUC using test-train approach

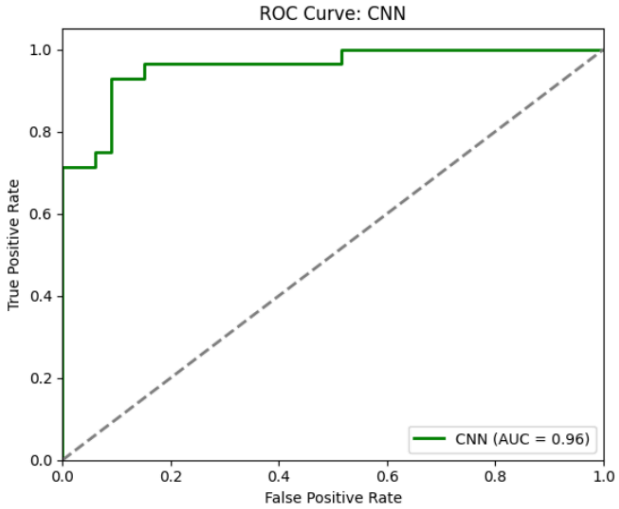


Figure 4. RNN models AUC using test-train approach

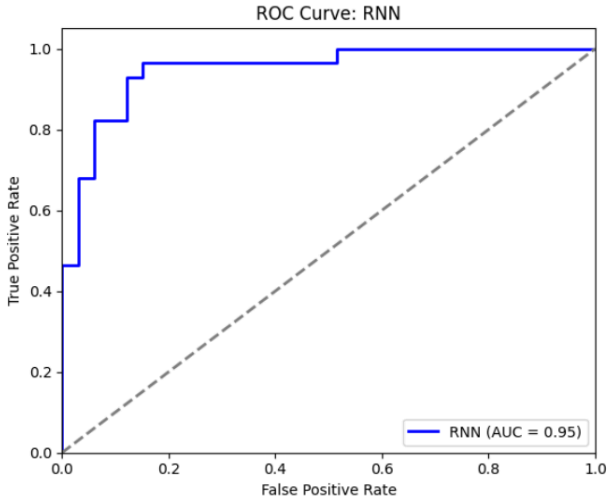


Figure 5. LSTM models AUC using test-train approach

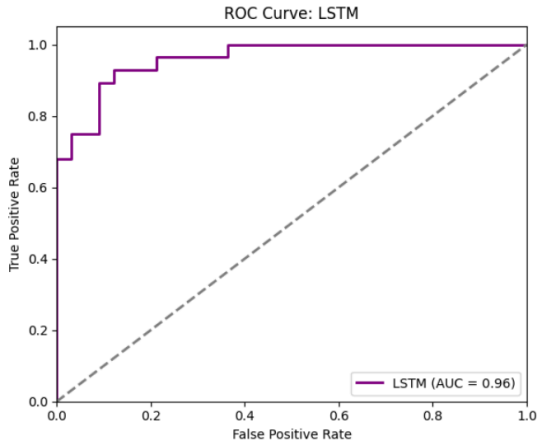


Figure 6. CNN-RNN models AUC using test-train approach

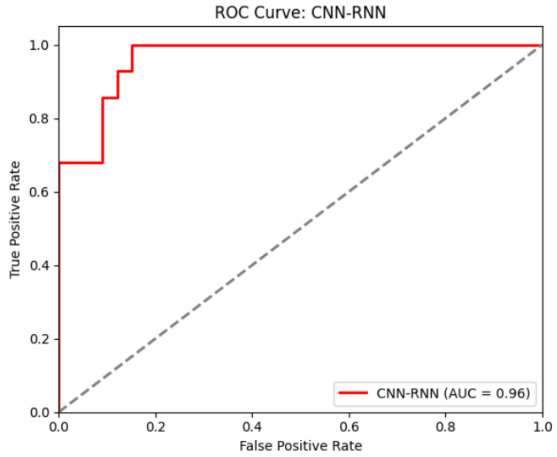


Figure 7. CNN-LSTM models AUC using test-train approach

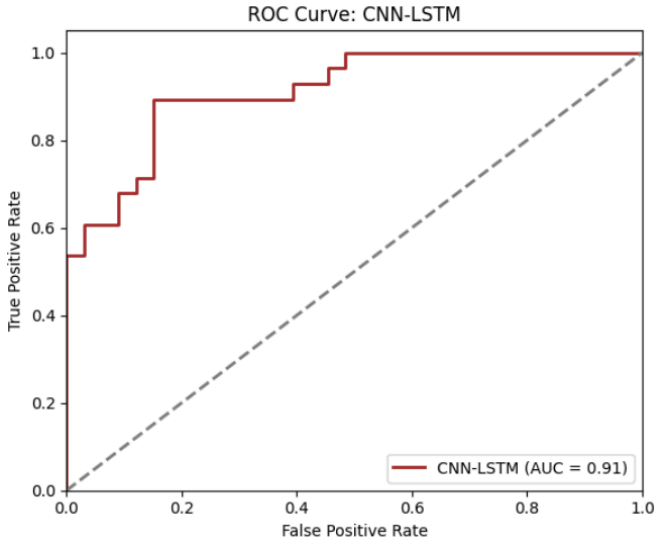
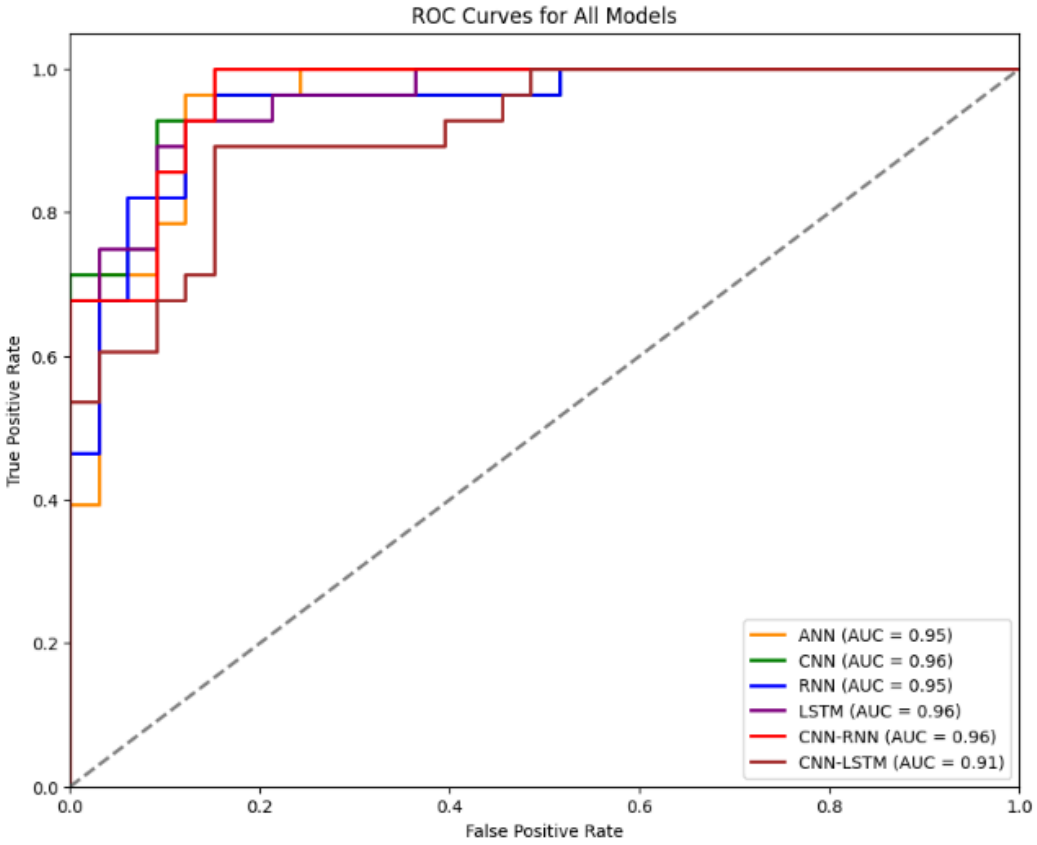


Figure 8. ROC Curves and AUC Values for All Models Using the Train-Test Split.



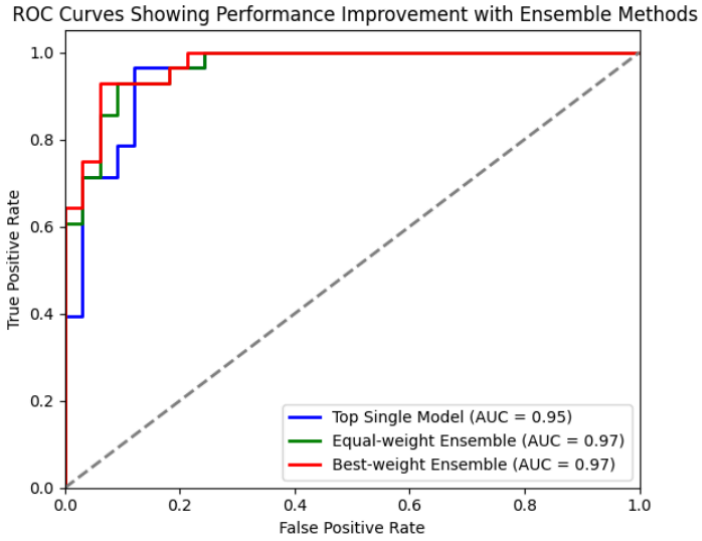
4.4 Ensemble Experiments

While individual deep learning models demonstrated strong performance in predicting heart disease, ensemble methods have often been shown to yield further improvements by leveraging the complementary strengths of multiple predictors. To explore this possibility, we implemented a simple ensemble strategy that combined the predicted probabilities from the top-performing models in our study—specifically, the ANN, RNN, and CNN models—using weighted averaging.

We first obtained probability estimates for the test set from each of the selected models. Rather than relying solely on binary predictions, we averaged their predicted probabilities. Initially, we assigned equal weights to each model's predictions, effectively computing a simple mean of their outputs. We then experimented with different weighting schemes to optimize the Area Under the ROC Curve (AUC), one of the key metrics in clinical predictive modeling.

The ensemble experiments yielded additional insights. As reported, the equal-weight ensemble of ANN, RNN, and CNN predictions achieved an accuracy of 0.9016, a precision of 0.8667, a recall of 0.9286, an F1-Score of 0.8966, and an AUC of 0.9665, with a log loss of 0.2911. Notably, this simple averaging approach already matched or slightly surpassed the performance metrics of the best individual models. Furthermore, by tuning the weights assigned to each model's probabilities, we identified a combination (0.5 for ANN, 0.5 for RNN, and 1.5 for CNN) that improved the AUC to 0.9719 and reduced the log loss to 0.2796, while maintaining the same accuracy and F1-score. These enhancements, although modest, confirm that ensembles can effectively leverage complementary modeling strengths of different architectures. This aligns with established ML research where ensembles often outperform single predictors, thus reinforcing the notion that even small gains in a clinical predictive context can be valuable. Such improvements may translate into more reliable decision support tools, guiding earlier and more accurate heart disease interventions.

Figure 9. AUC Curves for Ensemble Models



4.5 Discussion of Findings in Relation to Objectives

Recalling the objectives from Section 1.3, we aimed to (1) identify top-performing DL architectures, (2) ensure robust evaluation with multiple metrics, and (3) gain insights into feature importance and interpretability.

A. Identifying Top-Performing Architectures:
The ANN and RNN consistently produced excellent results, slightly outperforming or matching more complex architectures. While this might appear counterintuitive, it highlights that simpler, well-tuned architectures can be highly effective, particularly in scenarios where feature engineering and preprocessing reduce complexity.

B. Robust Evaluation and Metrics:
Employing accuracy, precision, recall, F1-score, AUC, and log loss offered a comprehensive understanding of each model's strengths and weaknesses. High AUC and favorable F1-scores confirm that these models are not only accurate but also sensitive to disease-positive cases—crucial in clinical decision-making.

C. Feature Importance and Interpretability:
Although this section primarily focuses on performance, the strong outcomes following feature selection support the idea that a carefully chosen subset of features can enhance both accuracy and interpretability. Future work will involve interpretability frameworks (e.g., SHAP, LRP) to elucidate which features drive predictions [21–23].

Figure 10. Training & Validation Accuracy/Loss Curves Over Epochs for ANN Model.

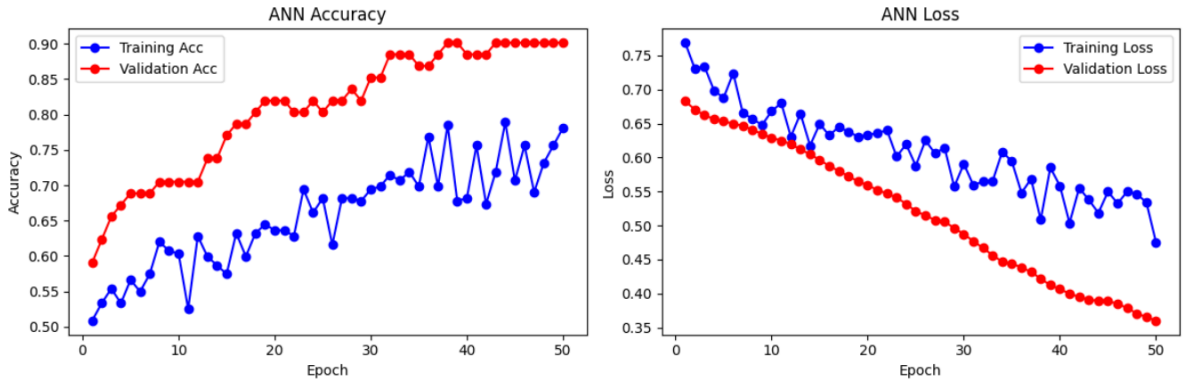


Figure 11. Training & Validation Accuracy/Loss Curves Over Epochs for CNN Model.

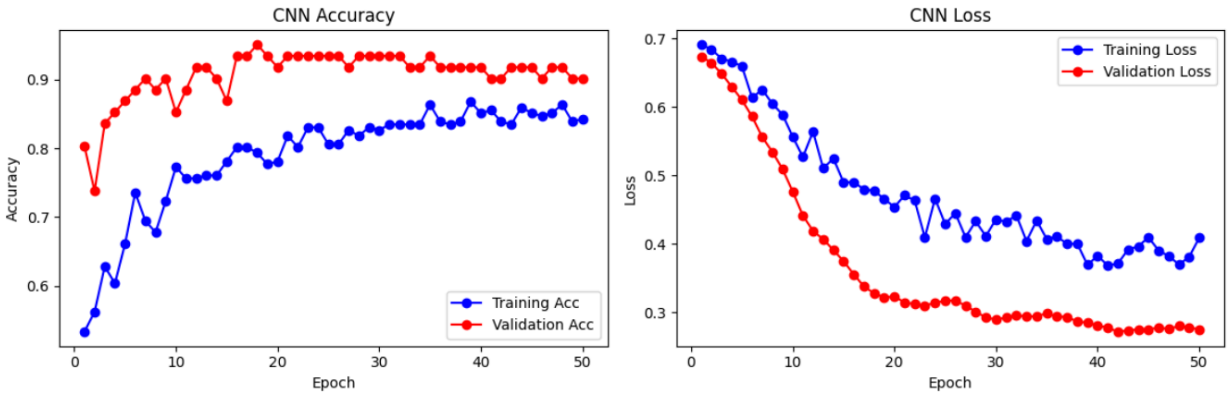


Figure 12. Training & Validation Accuracy/Loss Curves Over Epochs for the RNN Model.

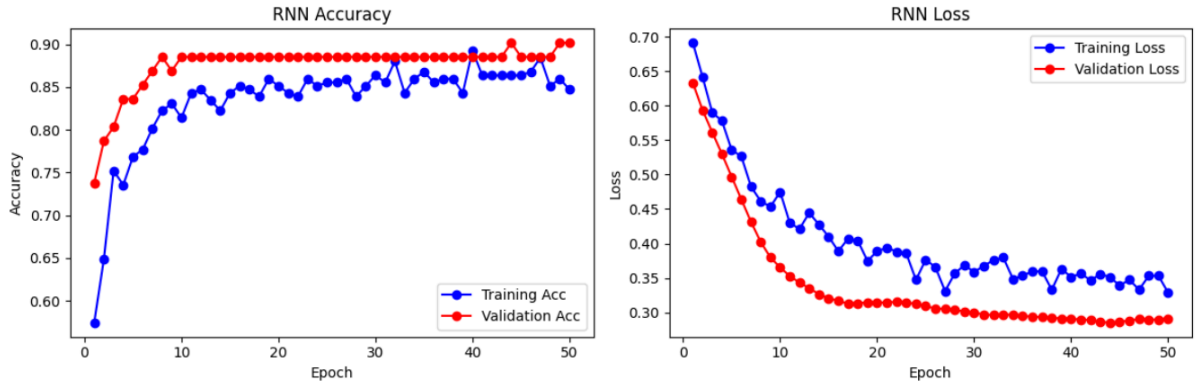


Figure 13. Training & Validation Accuracy/Loss Curves Over Epochs for the LSTM Model.

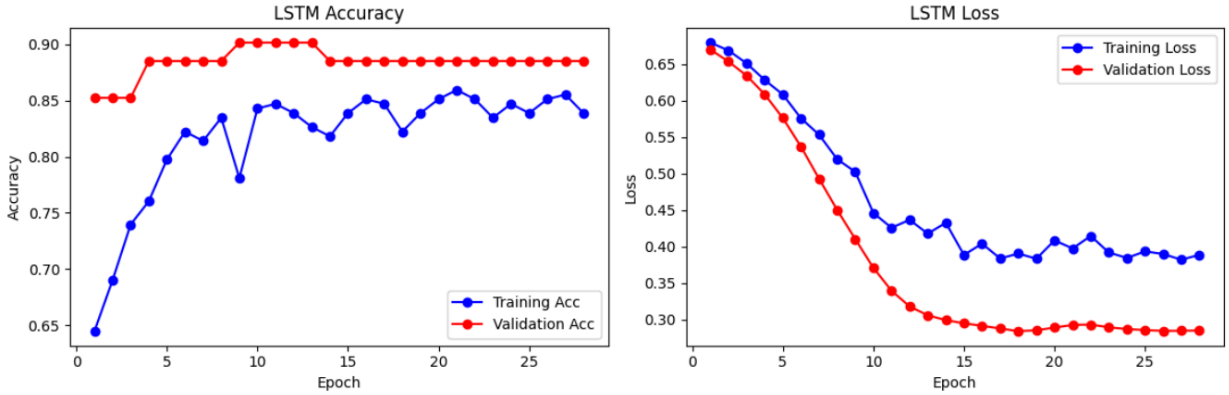


Figure 14. Training & Validation Accuracy/Loss Curves Over Epochs for the CNN-RNN Model.

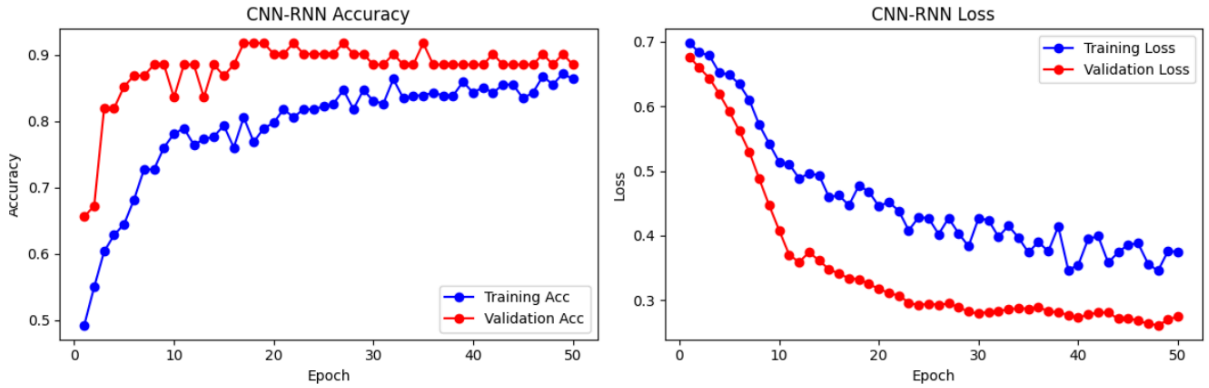
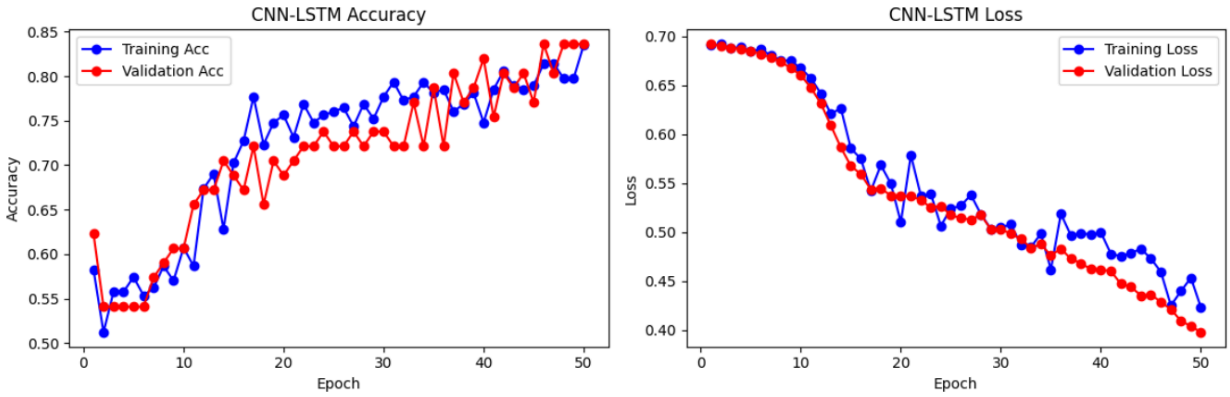


Figure 15. Training & Validation Accuracy/Loss Curves Over Epochs for the CNN-LSTM Model.



4.6 Comparisons with Existing Literature

Comparing our outcomes to previous studies (Section 2), which often reported accuracies ranging around 80–90% with single-model comparisons [6–10], our models achieved upper-tier performances. The ANN and RNN results align with earlier findings where neural networks surpassed traditional ML methods [16,17]. Our approach's novelty lies in systematically evaluating multiple DL architectures and confirming that simpler or moderately complex architectures can hold their own against more sophisticated hybrids under certain conditions.

In relation to interpretability and clinical trust, our study's emphasis on feature selection and performance metrics sets the stage for integrating model explanations—an advancement that many prior works have either not emphasized or only discussed superficially.

4.7 Implications and Limitations

The consistently high AUC and solid F1-scores are encouraging, suggesting that DL models can effectively aid in early heart disease risk detection. However, there are limitations to consider:

A. Dataset Size and Nature: The UCI Heart Disease Dataset, while standard and widely used, is relatively small and may not represent the full complexity and variability found in real-world clinical data. Larger, more heterogeneous datasets might yield different insights or amplify the benefits of more complex architectures.

B. Model Complexity vs. Simplicity: The finding that ANN and RNN models performed competitively suggests that complexity does not guarantee better performance. This might limit the generalization of this conclusion to other datasets or conditions where LSTM or CNN architectures might excel.

C. Interpretability Tools Not Fully Explored: While feature selection improved clarity, full deployment of interpretability methods remains a future direction. The ultimate clinical utility of these models will depend on explaining their predictions to clinicians and patients.

Comparison

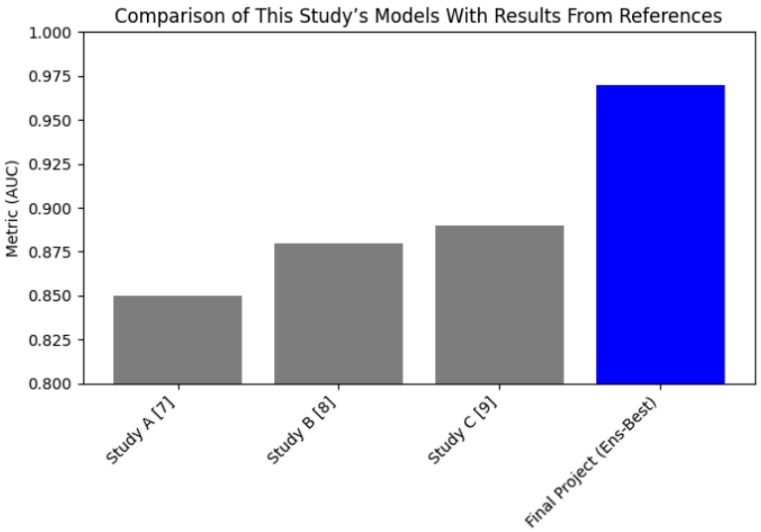
The primary objectives of this study were to systematically evaluate multiple deep learning (DL) architectures on the UCI Heart Disease Dataset and determine whether advanced or hybrid models confer tangible advantages over simpler baseline architectures. Additionally, the study aimed to situate these findings within the broader body of literature and explore ensemble strategies to further improve predictive performance.

5.1 Comparison with Previous Studies

Prior research applying traditional machine learning (ML) and early DL methods to the UCI Heart Disease Dataset often reported accuracies in the 80–90% range [6–10]. Classical ML models, including logistic regression, decision trees, and SVMs, generally achieved accuracies in the mid-80% range when carefully tuned [7–9]. Simple neural network approaches occasionally improved these outcomes, pushing performance into the upper 80s or low 90s [11,16–17]. These historical benchmarks suggest that substantial gains over standard ML techniques have proven challenging, prompting interest in more intricate or domain-specific architectures.

In our study, even relatively simple architectures—such as the ANN and RNN—consistently reached accuracies near 0.90 and AUC values exceeding 0.90 (Section 4). More complex models (CNN, LSTM, CNN-RNN, and CNN-LSTM) performed competitively but did not universally surpass these simpler configurations. This outcome highlights the crucial role of preprocessing, feature selection, and hyperparameter tuning. It also challenges the assumption that increasing architectural complexity necessarily yields performance improvements. The findings underscore the importance of a balanced approach that complements model sophistication with robust data preparation and selection of salient features.

Figure 16: Comparison of This Study's Models to Previously Reported Results in the Literature

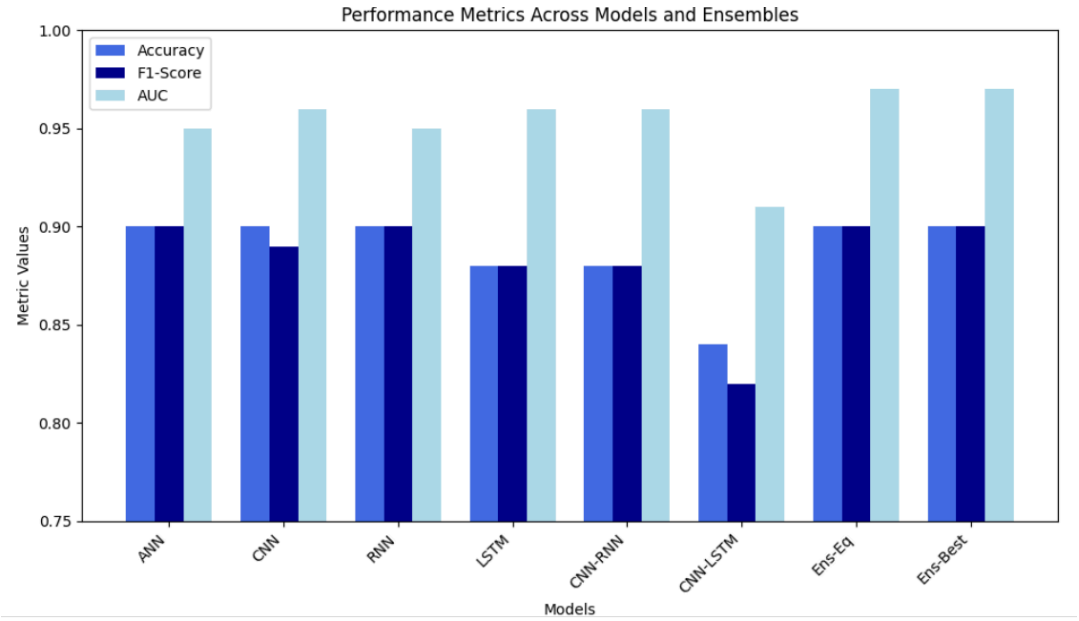


5.2 Ensemble Methods Compared to Individual Models

In extending our analysis to ensemble methods, we observed that combining predictions from multiple top-performing models (ANN, RNN, and CNN) elevated the performance slightly above the best single-model scores. The equal-weight ensemble achieved an AUC of 0.9665, and by tuning weights, we raised the AUC further to 0.9719. These ensemble results surpass both our best individual model performances and many historical benchmarks.

While the magnitude of improvement from ensembling was incremental, the significance lies in the consistency and robustness that ensembles often provide. Previous literature in ML and DL has repeatedly shown that even minor improvements in AUC or accuracy can be meaningful in a clinical environment where timely and accurate detection of heart disease risk factors can influence treatment outcomes [12–14]. In this regard, our ensemble findings align with broader ML practice, confirming that a combination of diverse models can enhance reliability and potentially lower variance.

Figure 16. Comparative Summary of Performance Metrics Across Models and Ensembles



5.3 Insights into Data and Feature Importance

The application of backward feature selection and normalization techniques appears to have contributed significantly to the strong baseline performance of even simple models. By ensuring that only the most informative features were retained, we may have alleviated the necessity for more complex architectures to extract relevant representations. This observation is consistent with some prior works where meticulous data preparation narrowed performance gaps between basic and advanced models [16–17].

As previously noted, interpretability remains an open avenue for future exploration. Studies have highlighted that applying post-hoc explainability methods like SHAP or LRP can reveal how models prioritize different clinical factors, providing additional clinical value [21–23]. The relatively high performance of simpler and ensemble models suggests that once interpretability frameworks are integrated, clinicians may glean meaningful insights into which

subsets of features (e.g., ST depression, maximum heart rate, or particular categorical indicators such as “thal”) are driving predictions.

5.4 Comparison of Validation Strategies and Metrics

We employed both an 80:20 train-test split and 10-fold cross-validation to ensure robustness (Sections 4.2 and 4.3). While cross-validation results showed slightly lower average metrics—an expected outcome due to more stringent testing across multiple folds—they confirmed the overall superiority of DL models over traditional benchmarks. This matches patterns reported in related literature, reinforcing the notion that the chosen evaluation methodologies are sound and that the models generalize adequately across multiple data partitions.

Considering a broader spectrum of metrics (accuracy, precision, recall, F1-score, AUC, and log loss) provided a more holistic interpretation of performance than accuracy alone. This multifaceted evaluation mirrors the recommendations of clinical ML guidelines, which emphasize the importance of understanding different error types and calibration levels, especially in high-stakes medical environments [2].

Conclusion

This research undertook a comprehensive evaluation of multiple DL architectures—encompassing ANN, CNN, RNN, LSTM, CNN-RNN, and CNN-LSTM models—applied to the UCI Heart Disease Dataset. Through rigorous preprocessing (including imputation of missing values, one-hot encoding of categorical features, normalization of continuous attributes, and backward feature selection), the study established a refined, high-quality dataset that minimized noise and emphasized salient risk factors [15–17]. Operating on this well-prepared dataset, even simpler DL models (e.g., ANN, RNN) attained accuracy scores around 0.90 and AUC values exceeding 0.90, thereby meeting or surpassing performance levels reported in previous literature [6–10].

Contrary to conventional expectations, more complex architectures did not universally outperform their simpler counterparts. While CNNs, LSTMs, and hybrids demonstrated robust performance, their advantages over the ANN and RNN models were not consistently substantial. This finding underscores the importance of domain-specific preprocessing, judicious feature selection, and careful hyperparameter tuning in unlocking model capabilities. Furthermore, the introduction of ensemble methods showed that combining model predictions could incrementally improve performance, reaching AUC values near 0.97.

Nonetheless, several limitations persist. The relatively small and homogeneous dataset may not encapsulate the complexity and diversity found in broader clinical populations. Future studies should evaluate these models against larger, heterogeneous patient cohorts and consider domain knowledge-based feature engineering. Moreover, while strong predictive metrics are essential, clinical applicability demands that models be interpretable and explainable. Incorporating techniques such as SHAP values or Layer-wise Relevance Propagation [21–23]

would provide insights into the clinical features driving predictions, bolstering trust and facilitating adoption in practice.

In summary, this investigation demonstrates that with rigorous preprocessing and informed feature selection, simpler DL models can perform competitively, challenging the notion that complexity is inherently superior. The potential for incremental gains through ensembling further positions these approaches as valuable tools in clinical decision support, aiding in earlier detection and more personalized management of heart disease. The findings pave the way for future work aimed at enhancing interpretability, scalability, and domain relevance, ultimately bridging the gap between state-of-the-art computational methods and tangible clinical outcomes.

Acknowledgments

The author extends gratitude to the Department of Mathematics at Arizona State University for their support during this research. The author also acknowledges the UCI Machine Learning Repository for providing access to the dataset used in this study.

Author Contributions

All of the contributions were made by: Ananya Bhargavi Kodali

Data Availability

The UCI Heart Disease Dataset used in this study is publicly available at the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml).

References

[1] World Health Organization. Cardiovascular diseases (CVDs). (2021). Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Benjamin EJ, Muntner P, Alonso A, et al. Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation. 2019;139(10):e56–e528.
[3] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–29.
[4] Shah RU, Abbasi SA, Attia ZI, et al. Artificial intelligence and machine learning in cardiovascular imaging: Opportunities and challenges. Curr Treat Options Cardiovasc Med. 2020;22(11):62.
[5] Dua D, Graff C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science; 2019.
[6] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 2002;26(1-2):1–24.
[7] Baxt WG. Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion. Neural Comput. 1990;2(4):480–489.

[8] Polat K, Güneş S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Comput Methods Programs Biomed. 2007;88(2):164–174.
[9] Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989;64(5):304–310.
[10] Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl. 2009;36(4):7675–7680.
[11] Yoon YJ, Cho S, Park HK. Efficient diagnosis system for congestive heart failure using heart rate variability and echocardiography. J Med Syst. 2013;37(2):9895.
[12] Zang X, Li X, Li L. A hybrid model based on CNN and XGBoost for predicting heart disease. In: Proceedings of the 2018 International Conference on Computer Science and Artificial Intelligence. ACM; 2018. p. 71–75.
[13] Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677; 2015.
[14] Singh N, Jindal H, Kumar P. Heart disease prediction system using LSTM recurrent neural network. Int J Innov Technol Exploring Eng. 2020;9(4):2354–2359.
[15] Jain YK, Bhandare SK. Min max normalization based data perturbation method for privacy protection. Int J Comput Commun Technol. 2011;2(8):45–50.
[16] Manjaiah D, Belete DM. Wrapper based feature selection techniques on EDHS-HIV/AIDS dataset. Eur J Mol Clin Med. 2020;7(8):2642–2657.
[17] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. Morgan Kaufmann. 2011;5(4):83–124.
[18] Wu J. Introduction to Convolutional Neural Networks. Natl Key Lab Novel Software Technol, Nanjing Univ. 2017.
[19] Sherstinsky A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D. 2020;404:132306.
[20] Xiao Y, Cho K. Efficient character-level document classification by combining convolution and recurrent layers. arXiv preprint arXiv:1602.00367; 2016.
[21] Rahman M, Islam D, Mukti RJ, Saha I. A deep learning approach based on convolutional LSTM for detecting diabetes. Comput Biol Chem. 2020;88:107329.
[22] Xie Y, Zhu C, Zhou W, Li Z, Liu X, Tu M. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. J Petrol Sci Eng. 2018;160:182–193.
[23] Chollet F. Keras. https://keras.io/; 2018 [Accessed: June 10, 2021].
[24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
[25] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: 3rd Int Conf for Learning Representations. 2014.
[26] Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315–1316.
[27] Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak. 2018;18(4):55–64.
[28] Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: Current methods and applications. BMC Med Res Methodol. 2017;17(1):1–19.
[29] Pierce R. Evaluating information: Validity, reliability, accuracy, triangulation. In: Research Methods in Politics: A Practical Guide. Sage Publications; 2008.
[30] Betechuoh BL, Marwala T, Tettey T. Autoencoder networks for HIV classification. Curr Sci. 2006;91(11):1467–1473.