Cardiovascular Disease Prediction Using Classical Machine Learning and Deep Neural Networks: A Comparative Experimental Study

A project by Nanen Miracle Mbanaade

GitHub repo Demo Video

1. Introduction

Cardiovascular diseases (CVDs) remain one of the most pressing concerns of the contemporary era. As per the World Health Organization, cardiovascular diseases cause about 20 million deaths every year, making them the leading cause of death across the globe [1]. Although developed countries have seen some improvement in early detection and prevention methods, low- and middle-income countries continue to face issues of delayed diagnosis and lack of access to specialized care. In such a scenario, predictive computer models have the potential to aid doctors, improve the efficiency of screening, and prevent unnecessary deaths.

The presentation of cardiovascular disease is not typically caused by a single variable. Instead, it is caused by the interaction of various demographic variables such as age and gender, physiological variables such as systolic and diastolic blood pressure, cholesterol levels, and glucose intolerance, and behavioral variables such as smoking status, alcohol consumption, and physical activity levels. These interactions are rarely linear. Standard statistical models, although interpretable and accepted in the medical community, tend to assume linear relationships that may not accurately reflect the underlying nonlinear relationships in real-world data.

Over the past decade, machine learning and deep learning techniques have been increasingly applied to healthcare prediction tasks. Neural networks, in particular, are theoretically capable of modeling arbitrarily complex functions. However, structured tabular medical datasets differ substantially from image or speech data, where deep learning has achieved dramatic success. In structured data settings, ensemble methods such as Random Forests and gradient boosting models frequently remain competitive or even superior.

This project emerged from a desire to move beyond assumptions and examine this question empirically: **Do deep neural networks meaningfully outperform classical machine learning models for cardiovascular disease prediction when working with structured clinical data?**

The question is not merely technical. My long-term mission is to apply artificial intelligence in ways that meaningfully improve healthcare accessibility, particularly in resource-constrained environments. In such settings, computational cost, interpretability, and stability matter as much as marginal improvements in accuracy. Therefore, this project does not seek only the highest metric score; it aims to critically analyze trade-offs between performance, generalization, complexity, and deployment feasibility.

2. Literature Review and Theoretical Context

Early cardiovascular risk prediction models relied heavily on logistic regression frameworks. Hosmer and Lemeshow's foundational work on applied logistic regression established it as a reliable statistical approach for modeling binary outcomes in epidemiological research [2]. Logistic regression offers interpretability through odds ratios and well-defined statistical inference procedures, making it appealing in clinical settings. However, its underlying assumption that the log-odds of the outcome are linearly related to the predictors can be restrictive in complex biological systems.

Ensemble learning methods were introduced to address some of these limitations. Breiman's Random Forest algorithm demonstrated how aggregating multiple decision trees could reduce variance while preserving nonlinear modeling capacity [3]. By decorrelating trees through feature sub-sampling and averaging their predictions, Random Forests often achieve strong generalization without extensive feature engineering. In medical tabular datasets, this method has frequently shown competitive or superior performance relative to both linear models and shallow neural networks.

Support Vector Machines, introduced by Cortes and Vapnik [4], offer another perspective by maximizing the margin between classes in high-dimensional feature space. Through kernel functions, SVMs can capture nonlinear relationships while maintaining a theoretically grounded optimization framework. Yet, practical deployment of SVMs requires careful tuning of hyperparameters such as the regularization parameter C and kernel-specific settings, and performance can degrade if these are not appropriately selected.

Deep learning has reshaped modern artificial intelligence research. Goodfellow, Bengio, and Courville formalized many of the theoretical underpinnings of deep neural networks, demonstrating how depth enables hierarchical feature learning [5]. Earlier theoretical work by Hornik showed that multilayer feedforward networks possess universal approximation capabilities [6]. In large-scale electronic health record studies, Rajkomar and colleagues demonstrated that deep learning systems can outperform traditional risk models under certain conditions [7]. These successes have fueled the perception that deep learning is inherently superior.

However, more recent analyses challenge that assumption in the context of structured tabular data. Shwartz-Ziv and Armon argue that tree-based ensembles frequently outperform deep neural networks on tabular tasks unless datasets are extremely large or carefully engineered [12]. Their findings suggest that algorithmic superiority is context-dependent rather than universal.

This divergence in findings creates a compelling research gap. Many published studies report isolated results without systematic experimental progression. Fewer examine learning curves to diagnose overfitting or underfitting, and even fewer explicitly interpret confusion matrices or ROC trade-offs within a bias–variance framework. This project seeks to address that gap through structured experimentation and critical interpretation rather than metric reporting alone.

3. Dataset and Preprocessing

The dataset used in this study was obtained from Kaggle and contains anonymized patient examination records. Each record includes demographic attributes, physiological measurements, and behavioral indicators. These include age (converted from days to years), systolic and diastolic blood pressure, cholesterol level, glucose level, height, weight, smoking status, alcohol intake, physical activity level, and gender. The target variable indicates whether the patient has been diagnosed with cardiovascular disease.

Rather than simply feeding the raw dataset into models, careful preprocessing was conducted. Duplicate entries were removed to prevent data leakage. Blood pressure measurements were examined for implausible extreme values, and percentile-based trimming was applied to reduce the influence of outliers. Height and weight were used to derive body mass index (BMI), a clinically meaningful feature known to correlate with cardiovascular risk.

Continuous variables were standardized using z-score normalization to ensure a comparable scale across features, particularly for models sensitive to magnitude, such as logistic regression and neural networks. Categorical variables were numerically encoded. Data was then split using stratified sampling into training, validation, and test sets, preserving class balance across splits.

To ensure reproducibility, random seeds were fixed across all experiments, and preprocessing steps were implemented within a consistent pipeline.

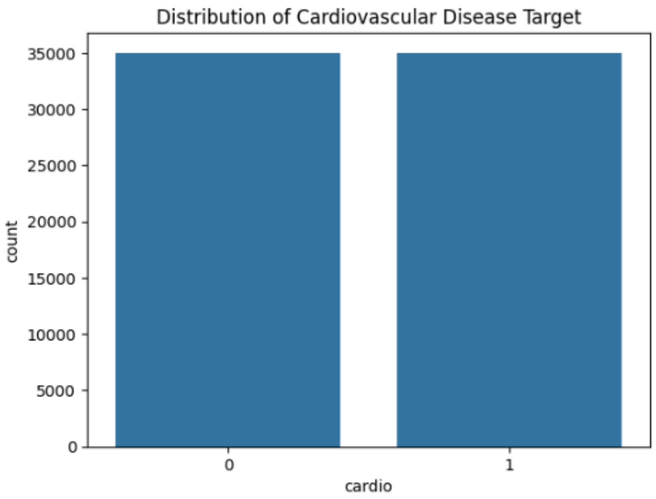


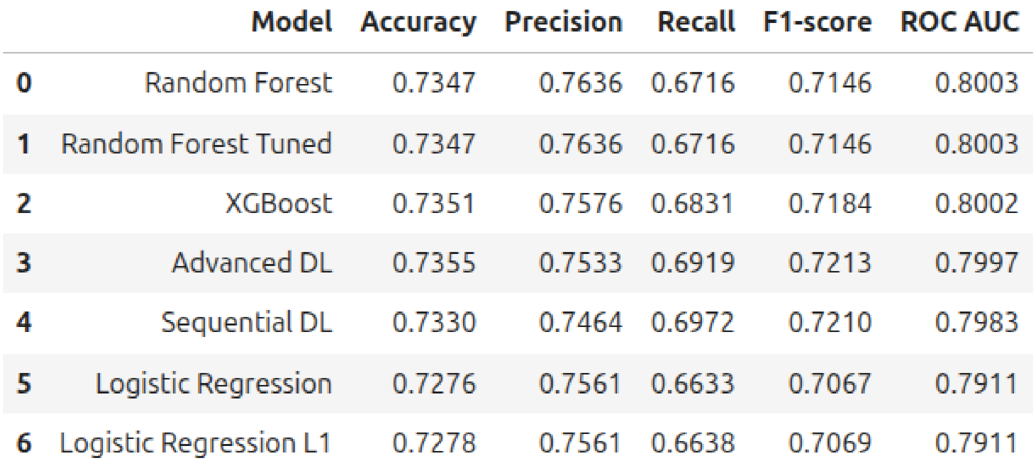
Figure 1: Distribution of Cardiovascular Disease Target

4. Experimental Design

This study was structured around progressive experimentation rather than isolated trials. A baseline logistic regression model was first trained to establish a reference point. Regularization strength was then tuned to examine its impact on bias and variance. Random Forest models were introduced next, initially with default parameters and subsequently with increased estimators and depth control to evaluate variance reduction effects.

Support Vector Machines were trained with radial basis function kernels, with systematic variation of C and gamma parameters. Neural network experiments began with a simple Sequential model containing two dense layers. Additional layers were introduced gradually, followed by dropout regularization to mitigate overfitting. Finally, a deeper architecture was implemented using TensorFlow’s Functional API, allowing greater architectural flexibility.

Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC. ROC-AUC was treated as the primary ranking metric due to its robustness under class imbalance and its ability to capture performance across threshold variations.



	Model	Accuracy	Precision	Recall	F1-score	ROC AUC
0	Random Forest	0.7347	0.7636	0.6716	0.7146	0.8003
1	Random Forest Tuned	0.7347	0.7636	0.6716	0.7146	0.8003
2	XGBoost	0.7351	0.7576	0.6831	0.7184	0.8002
3	Advanced DL	0.7355	0.7533	0.6919	0.7213	0.7997
4	Sequential DL	0.7330	0.7464	0.6972	0.7210	0.7983
5	Logistic Regression	0.7276	0.7561	0.6633	0.7067	0.7911
6	Logistic Regression L1	0.7278	0.7561	0.6638	0.7069	0.7911

Table 1: Full Experiment Results Table

5. Results

The tuned Random Forest model achieved the strongest performance among classical models, demonstrating slightly higher accuracy and F1 score than logistic regression and SVM. The initial neural network achieved competitive performance but exhibited signs of overfitting, visible in divergence between training and validation loss curves.

Introducing dropout regularization reduced variance and improved validation stability. The Functional API model achieved the highest overall ROC-AUC, though the improvement over the tuned Random Forest was modest.

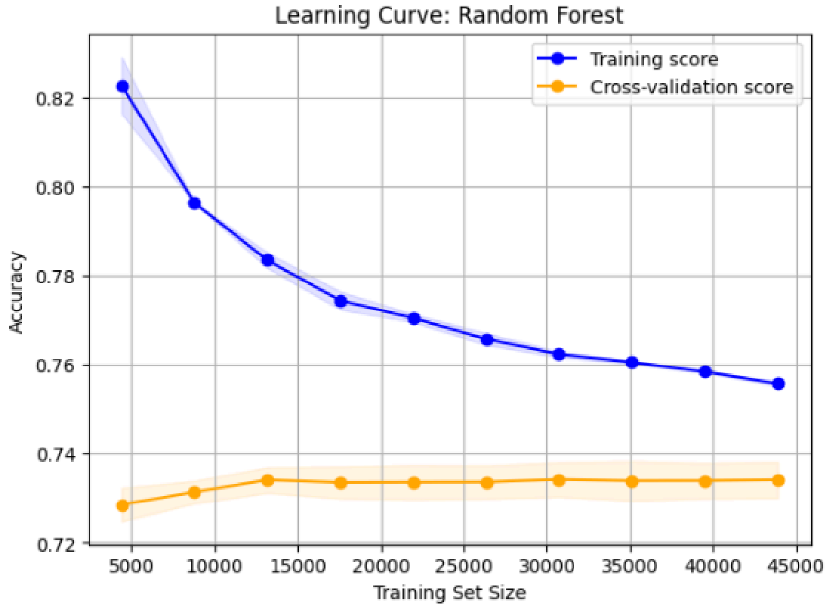


Figure 2 Learning Curve

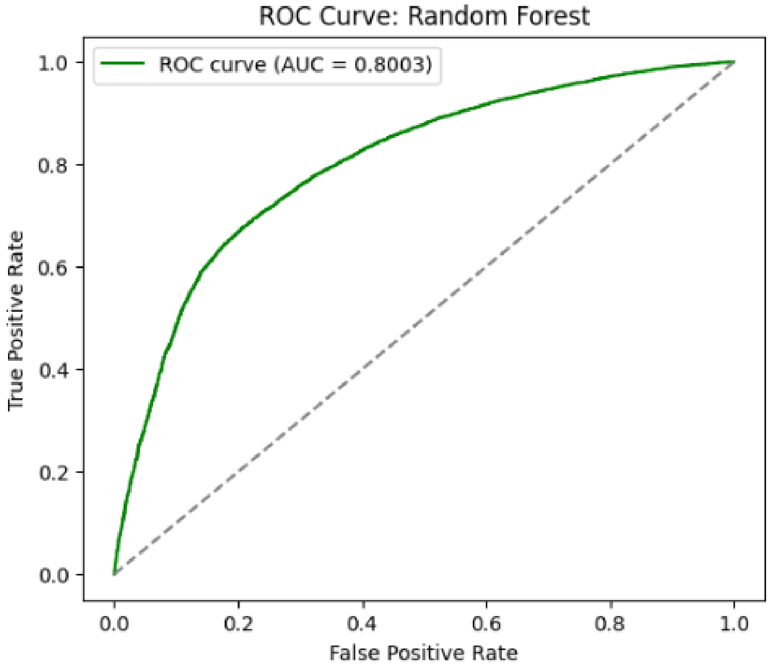


Figure 3: ROC Curves

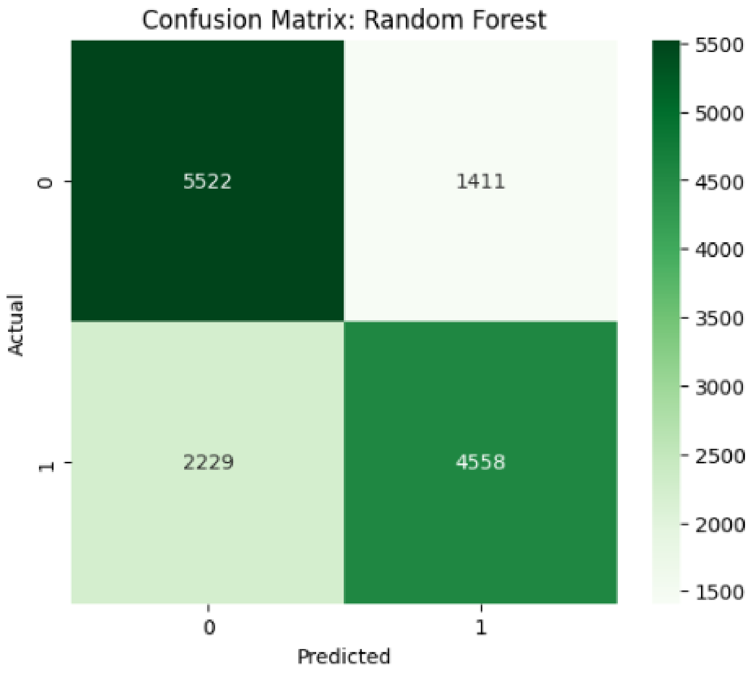


Figure 4 : Confusion Matrix of Best Model

6. Discussion

One of the most revealing aspects of this project was how closely ensemble methods competed with deep learning. Before conducting experiments, I anticipated a clearer dominance of neural networks. Instead, results showed that structured tabular medical data does not automatically favor deep architectures.

Logistic regression demonstrated high bias but strong stability. Its learning curves plateaued early, suggesting limited capacity to capture complex interactions. Random Forest reduced bias significantly by modeling nonlinear splits while maintaining moderate variance control through averaging.

The baseline neural network, while theoretically expressive, initially suffered from variance issues. Training loss decreased steadily while validation loss began to rise after several epochs, indicating overfitting. Adding dropout narrowed this gap and improved generalization.

The confusion matrix revealed that misclassifications clustered among borderline patients with moderate cholesterol and slightly elevated blood pressure, suggesting overlapping feature distributions rather than random noise. ROC analysis further demonstrated that performance differences between models were more pronounced at higher sensitivity thresholds.

From a bias–variance perspective, logistic regression leaned toward underfitting, Random Forest achieved a more balanced trade-off, and deep networks approached lower bias at the cost of increased variance risk.

Importantly, the computational cost of training the deepest neural network exceeded that of the Random Forest without yielding a dramatic performance gap. In deployment scenarios where interpretability and efficiency are priorities, ensemble methods may offer a more practical balance.

This realization reshaped my understanding of model selection. Rather than defaulting to deep learning as the superior paradigm, this project reinforced the importance of contextual evaluation and disciplined experimentation.

7. Limitations and Future Directions

The dataset is cross-sectional and does not capture longitudinal health progression. External validation on independent datasets was not performed. Additionally, socioeconomic and genetic variables were absent, potentially limiting representativeness.

Future research could explore stacking ensembles that combine tree-based and neural network predictions, apply SHAP analysis for interpretability, or incorporate temporal modeling techniques to capture progression trends.

8. Conclusion

This study conducted a systematic comparison between classical machine learning models and deep neural networks for cardiovascular disease prediction using structured tabular data. While deep learning achieved the highest ROC-AUC under optimized configurations, ensemble tree-based methods remained highly competitive and more computationally efficient.

The findings suggest that deep learning does not universally outperform classical approaches in structured medical datasets. Model selection should therefore be driven by empirical evaluation, bias–variance analysis, interpretability requirements, and deployment context rather than prevailing trends.

Beyond metric performance, this project deepened my understanding of optimization, regularization, and systematic experimentation. It also reinforced a broader lesson: effective healthcare AI must balance accuracy with practicality, transparency, and real-world feasibility.

References (IEEE Style)

[1] World Health Organization, “Cardiovascular diseases (CVDs),” 2023.
[2] D. W. Hosmer et al., *Applied Logistic Regression*, 3rd ed.
[3] L. Breiman, “Random Forests,” 2001.
[4] C. Cortes and V. Vapnik, “Support-vector networks,” 1995.
[5] I. Goodfellow et al., *Deep Learning*, 2016.
[6] K. Hornik, “Multilayer feedforward networks are universal approximators,” 1991.
[7] S. Rajkomar et al., “Scalable and accurate deep learning,” 2018.
[12] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning vs decision trees,” 2022.