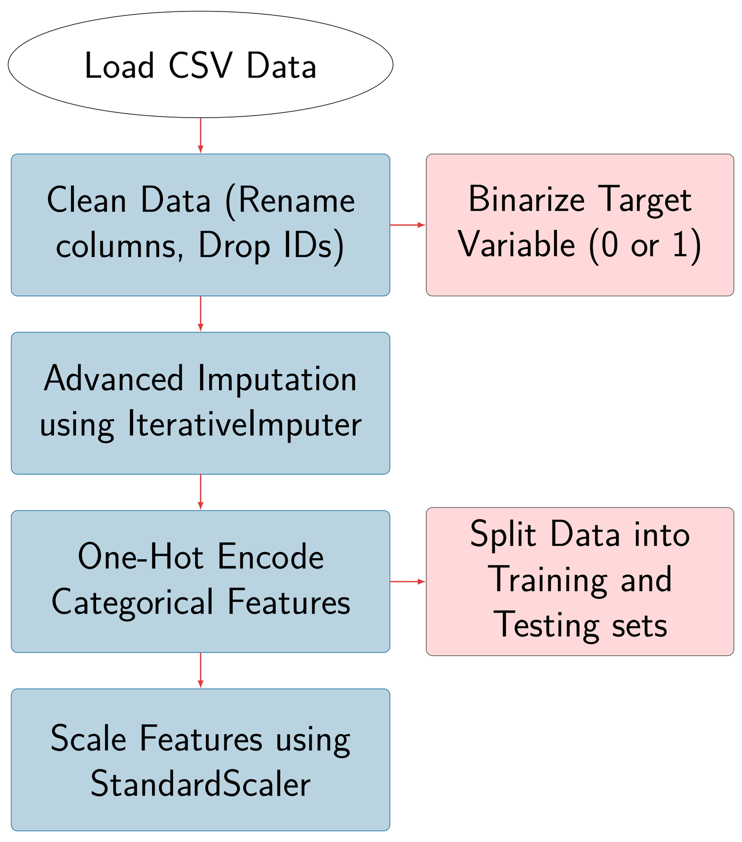
Comparative Study of Heart Disease Prediction Models

Liyans Mathews (24201531), Aleena Jacob (24210559)

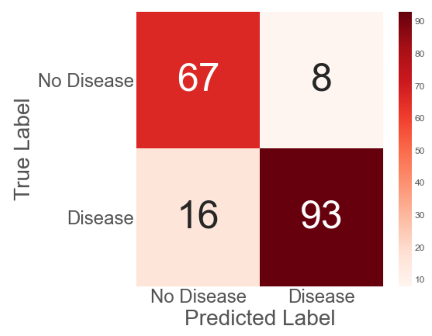
Introduction

- **Dataset:** Heart Disease UCI Cleveland dataset
- **Aim:** To identify the optimal ML model for predicting heart disease by comparing the performance of different models

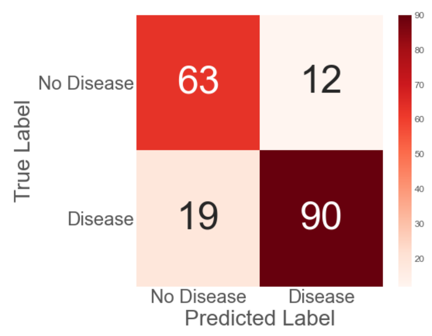
Pre-Processing Steps



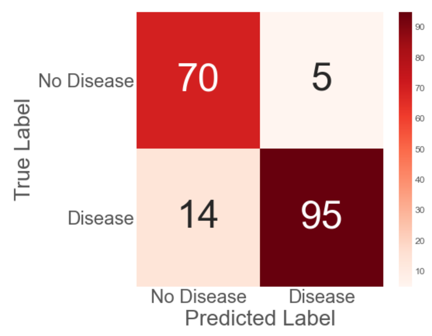
Confusion Matrix



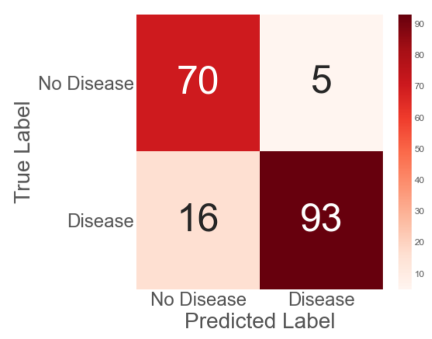
(a) Logistic Regression



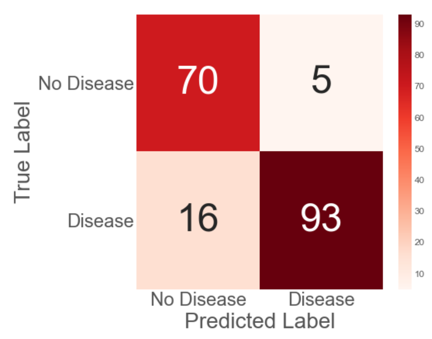
(b) Naive Bayes



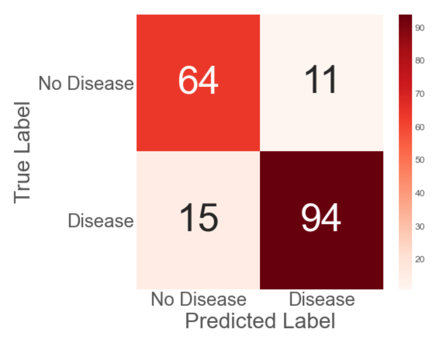
(c) Random Forest



(d) SVM



(e) XGBoost



(f) DenseNet

Figure 1: These confusion matrices illustrate the classification performance of each model on heart disease prediction. The Random Forest matrix shows the fewest misclassifications, indicating stronger predictive accuracy compared to other models.

Model Performance Metrics

Model	Accuracy	ROC AUC	PR AUC	F1-Score	Sensitivity	Specificity
Random Forest	0.897	0.961	0.972	0.909	0.872	0.933
Support Vector Machine	0.886	0.943	0.968	0.899	0.853	0.933
XGBoost	0.886	0.945	0.965	0.899	0.853	0.933
Logistic Regression	0.870	0.929	0.952	0.886	0.853	0.893
Autoencoder + DenseNet	0.848	0.922	0.938	0.868	0.844	0.853
Naive Bayes	0.832	0.912	0.931	0.853	0.826	0.840

Table 1: Performance metrics of all evaluated models.

Sensitivity vs Specificity

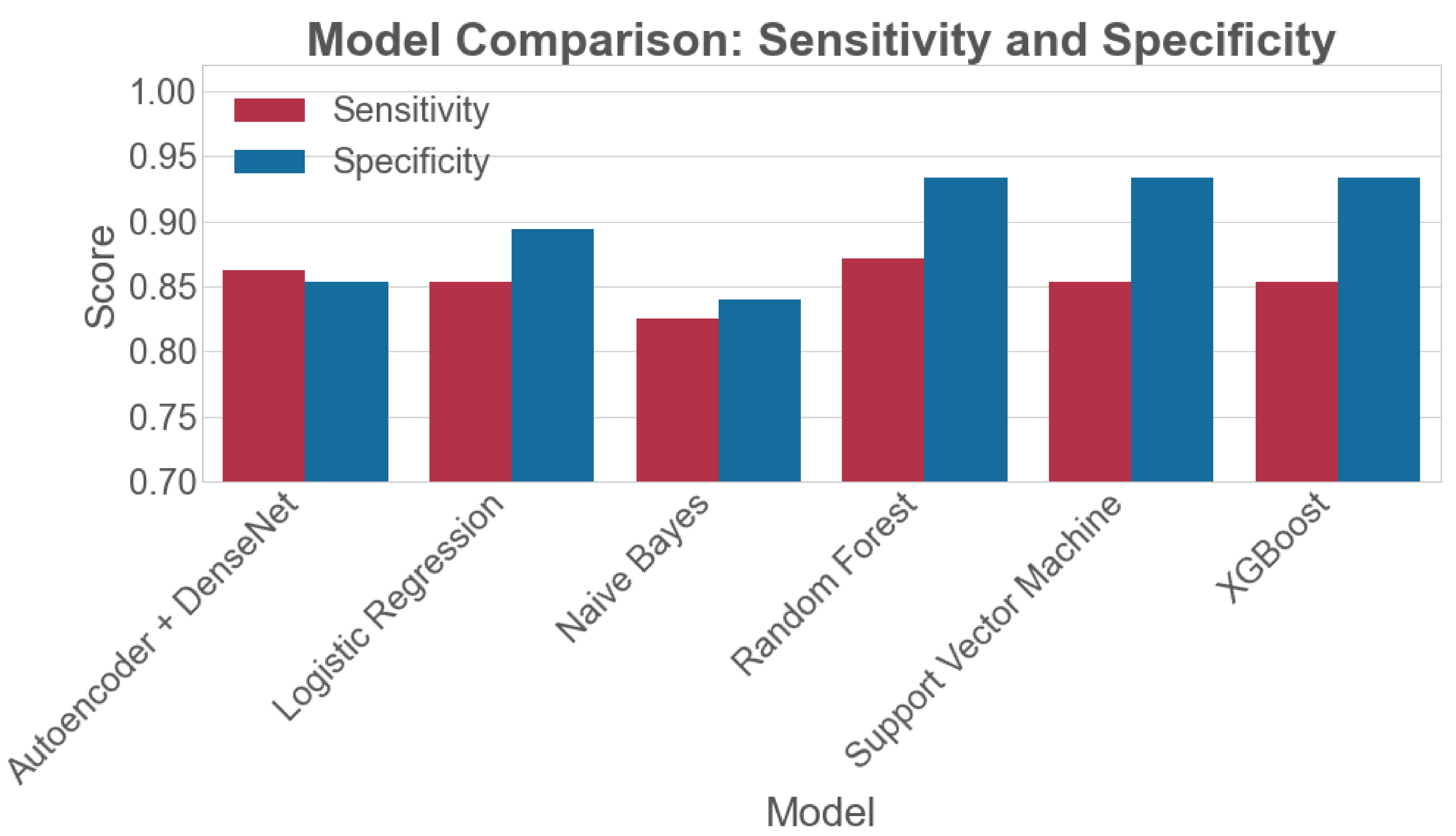


Figure 2: This plot compares the true positive rate (sensitivity) and true negative rate (specificity) for all models. Random Forest achieves the best balance, demonstrating its effectiveness in correctly identifying both patients with and without heart disease.

ROC Curve

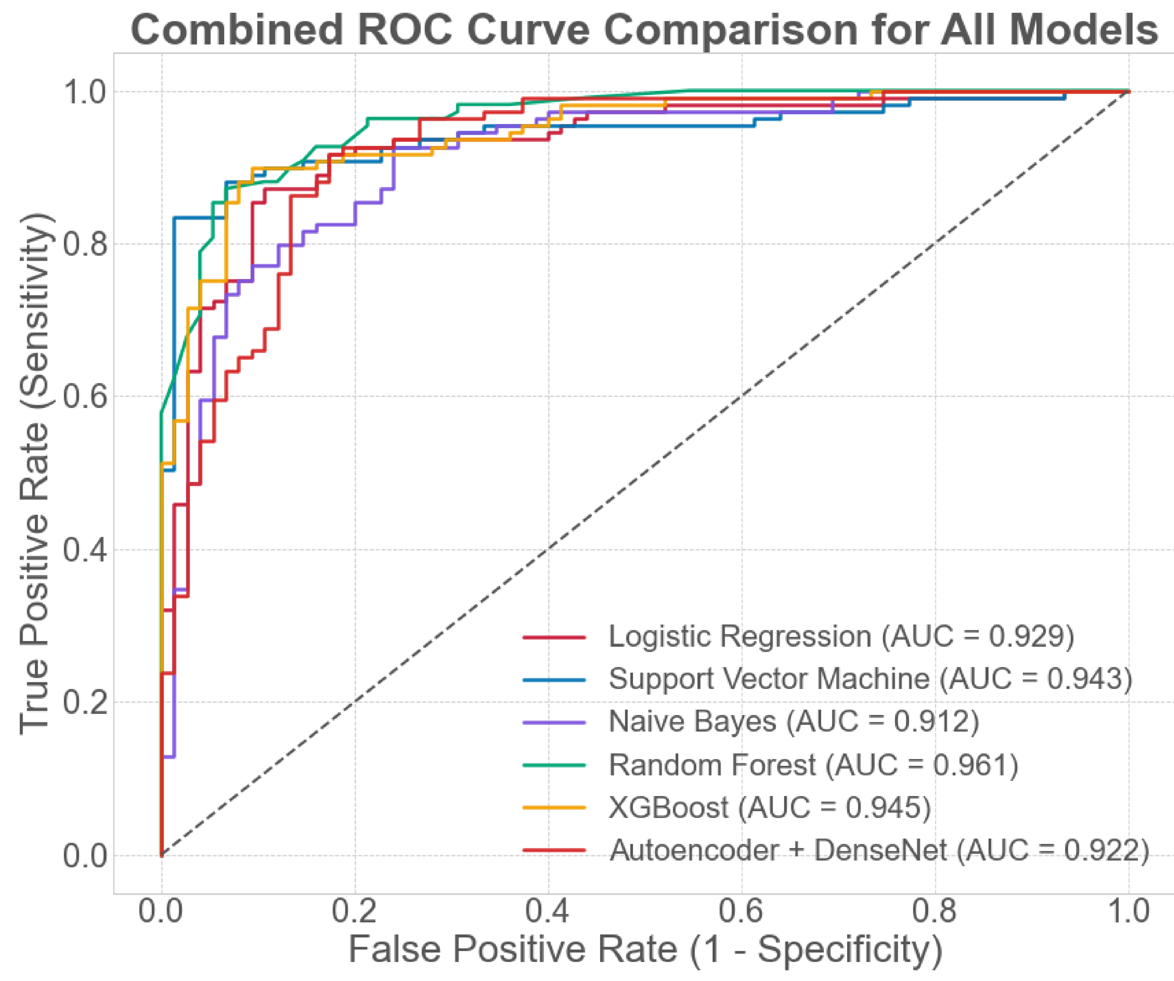


Figure 3: The Receiver Operating Characteristic (ROC) curves highlight each model's ability to distinguish between positive and negative cases. Random Forest has the highest Area Under the Curve (AUC) of 0.961, indicating superior discriminative performance over other models.

Model Accuracy and F1 Score Comparison

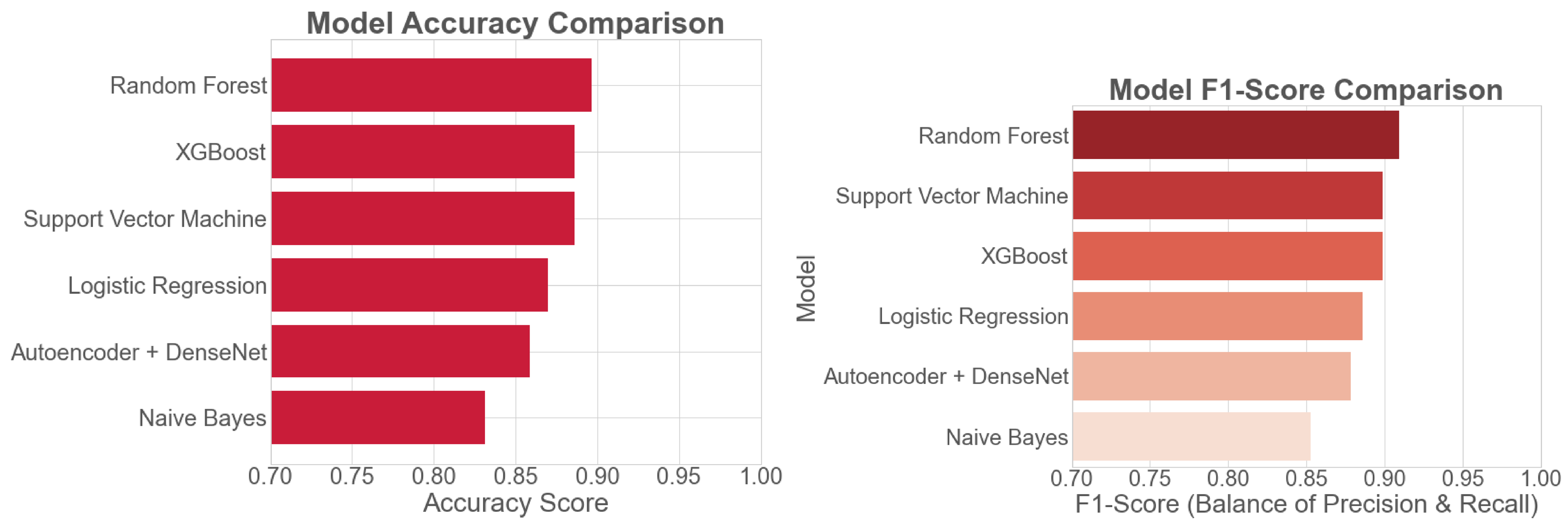


Figure 4: This plot compares the overall accuracy and F1-Score for each model. Random Forest again leads in both metrics, showing it balances precision and recall effectively, which is critical for reliable clinical predictions.

Precision Recall Curve

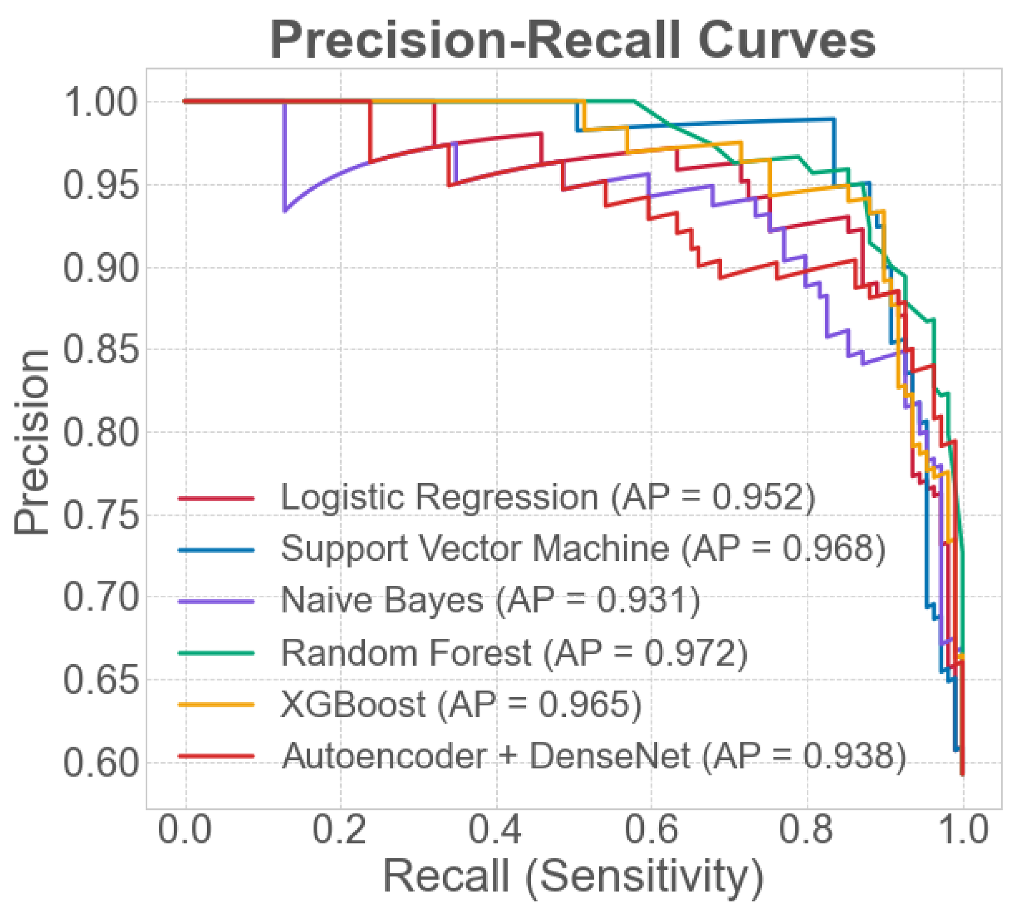


Figure 5: The Precision-Recall curves illustrate the trade-off between precision and recall for all models. The Random Forest model consistently maintains higher precision and recall, reflecting its robustness in predicting heart disease cases.

Conclusion

Across all evaluated performance metrics, Random Forest consistently outperforms the other models. Its superior results can be attributed to its ability to capture complex, non-linear relationships and interactions between clinical features in the UCI Heart Disease dataset, while also being robust to noise and overfitting through the use of multiple decision trees.

Random forest works because:

- **Ensemble Learning:** Combines multiple decision trees trained on random subsets, reducing overfitting.
- **Captures Non-Linear Relationships:** Models complex interactions among clinical features effectively.
- **Noise Resistance:** Averaging many trees reduces sensitivity to noise and outliers.
- **Balances Bias and Variance:** Provides accurate and generalizable predictions.
- **Feature Importance:** Identifies key predictors, improving interpretability and performance.

References

1. Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. *PMC*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11929227/
2. Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.
3. Alghamdi, N. S., Zakariah, M., Shankar, A., Viriyasitavat, W. (2024). Heart disease prediction using autoencoder and DenseNet architecture. *Egyptian Informatics Journal, 28*, 100559.