DOI: http://dx.doi.org/10.26483/ijarcs.v16i3.7280

Volume 16, No. 3, May-June 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTING HEART DISEASE

Dr. Bhanu Prakash Paruchuri
Information Technology (IT)
University of the Cumberlands (UOTC)
Williamsburg, Kentucky, USA

Abstract: Cardiovascular diseases, particularly heart disease, remain among the leading causes of morbidity and mortality globally [1]. As the prevalence of heart-related conditions increases, there is a growing demand for early diagnostic systems that can support clinical decision-making and preventive care [2]. This whitepaper presents a comparative analysis of multiple machine learning algorithms applied to the UCI Heart Disease dataset [3], leveraging both statistical and predictive modeling approaches to identify patterns associated with heart disease. The study begins with an in-depth exploratory data analysis (EDA) to uncover trends, outliers, and correlations among clinical attributes such as age, cholesterol levels, resting blood pressure, and electrocardiographic results [4]. Following EDA, a suite of machine learning models—including Logistic Regression, Random Forest, and Gradient Boosting—are implemented to classify patients based on the likelihood of heart disease presence. Each model is evaluated using robust metrics including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), enabling a performance-driven comparison [5]. Our findings indicate that ensemble-based models such as Gradient Boosting and Random Forest consistently outperform baseline models in predictive accuracy and sensitivity, making them ideal candidates for integration into clinical diagnostic tools [6]. The insights from this study highlight the critical role of feature selection, preprocessing, and model interpretability in healthcare AI applications [7]. This work contributes to the ongoing advancement of data-driven health technologies by demonstrating the potential of machine learning in enhancing the early detection and risk stratification of heart disease.
In the broader context, the United States' healthcare expenditure underscores the urgency for efficient diagnostic tools. In 2023, U.S. healthcare spending reached $4.9 trillion, accounting for 17.6% of the nation's Gross Domestic Product (GDP), with per capita spending at $14,570 [8]. Notably, hospital care, physician and clinical services, and retail prescription drugs collectively accounted for 60% of total spending [9]. Despite this substantial investment, the U.S. continues to face challenges in achieving optimal health outcomes, particularly in managing chronic diseases like heart disease. Implementing effective machine learning models for early detection can play a pivotal role in improving patient outcomes and reducing healthcare costs [10].

Keywords: AI, Machine Learning, Heart failure, Prediction, Diabetes, Random Forest Algorithm, lifestyle, Early Detection Algorithms

1. INTRODUCTION

Cardiovascular disease (CVD), particularly heart disease, remains the single largest cause of death globally, responsible for approximately 17.9 million deaths each year according to the World Health Organization (WHO) [11]. In the United States, heart disease alone accounts for nearly 700,000 deaths annually, representing about 1 in every 5 deaths [12]. The burden of this disease is not only measured in human lives but also in the economic toll it imposes on patients, caregivers, and healthcare systems. From frequent hospitalizations and diagnostic tests to surgical interventions and long-term medication, the management of heart disease contributes significantly to national healthcare expenditures [13].

In 2023, healthcare spending in the United States reached $4.9 trillion—approximately 17.6% of the Gross Domestic Product (GDP)—with a substantial portion directed toward managing chronic conditions like heart disease, diabetes, and hypertension [14]. Despite these investments, the nation continues to face major challenges in ensuring timely diagnosis, equitable access to care, and efficient use of healthcare resources. This scenario underscores the critical need for scalable, intelligent tools that can assist clinicians in identifying high-risk patients early and guiding them toward appropriate care pathways [15].

Recent advances in artificial intelligence (AI) and machine learning (ML) have opened new frontiers in medical diagnostics and predictive analytics. The practical application of predictive data analytics is vital for creating cost-effective healthcare strategies focused on managing chronic conditions like lung cancer [16]. By leveraging data-driven insights, healthcare organizations can improve care quality, enhance patient outcomes, reduce prevalence rates, and decrease healthcare costs [17]. ML algorithms can process vast and complex datasets, identify hidden patterns, and make data-driven predictions with high accuracy—capabilities that are especially beneficial in cardiology, where disease symptoms often overlap and evolve subtly over time [18]. However, the practical implementation of ML in clinical settings is still emerging, necessitating robust comparative studies that evaluate algorithm performance on real-world datasets [19].

This whitepaper addresses this need by presenting a comparative analysis of machine learning models for predicting heart disease using the publicly available UCI Heart Disease dataset. The dataset comprises anonymized clinical records for 303 patients, with features including age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiogram results, maximum heart rate, and ST depression [20]. These features are commonly collected during routine cardiac evaluations,

making the dataset a suitable proxy for real-world screening scenarios.

The methodology involves a step-by-step workflow beginning with data cleaning and exploratory data analysis (EDA) to understand underlying feature distributions and correlations. We apply various supervised machine learning algorithms—including Logistic Regression, Random Forest, and Gradient Boosting—to build predictive models. These models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to identify which algorithm offers the best trade-off between sensitivity and specificity in detecting heart disease [21].

In addition to model performance, the whitepaper explores:

- Feature importance and interpretability: Identifying which clinical variables most influence the model's decisions.
- Bias and generalizability: Considering how well the models perform across different subgroups (e.g., age, gender).
- Scalability and integration potential: Assessing how these models can be integrated into electronic health record (EHR) systems or clinical decision support tools. However, need to consider the challenges in EHR usage, the potential of AI solutions, successful case studies, and the ethical considerations and regulatory frameworks necessary for a smooth implementation of AI in healthcare settings [22].

Limitations and ethical concerns: Discussing data limitations, transparency, and responsible AI deployment in clinical environments.

The ultimate goal is not merely to build accurate models, but to demonstrate how data science can play a transformative role in preventive cardiology. By enabling early identification of at-risk individuals, these models have the potential to support proactive interventions, reduce hospital readmissions, and improve overall population health outcomes. Furthermore, this work contributes to global health objectives, including those outlined by the United Nations Sustainable Development Goal (SDG) 3: Ensure healthy lives and promote well-being for all at all ages. In particular, Target 3.4 aims to reduce premature mortality from non-communicable diseases, including CVD, by one-third by 2030 through prevention and treatment.

In summary, this whitepaper provides a rigorous yet practical framework for applying machine learning to one of the most pressing healthcare challenges of our time. It bridges the gap between data science research and clinical application, offering insights that are valuable to researchers, healthcare providers, policymakers, and technology developers alike.

2. METHODOLOGY

This section outlines the step-by-step process followed to develop, train, evaluate, and compare multiple machine learning models for heart disease prediction. The methodology consists of six major phases: data acquisition, preprocessing, exploratory data analysis (EDA), model development, performance evaluation, and comparative analysis.

I. Data Acquisition: The study utilizes the publicly available UCI Heart Disease dataset, a benchmark dataset in medical machine learning research [23]. It contains 303 patient records, each with 14 attributes including demographic, clinical, and exercise-induced variables. The target variable is binary, indicating the presence (`1`) or absence (`0`) of heart disease.
Key features include:
* Age, sex
* Chest pain type (cp)
* Resting blood pressure (trestbps)
* Serum cholesterol (chol)
* Fasting blood sugar (fbs)
* Resting electrocardiographic results (restecg)
* Maximum heart rate achieved (thalach)
* Exercise-induced angina (exang)
* ST depression (oldpeak)
* Slope of the peak exercise ST segment (slope)
* Number of major vessels colored by fluoroscopy (ca)
* Thalassemia condition (thal)

II. Data Preprocessing: Preprocessing is crucial to ensure data quality and model reliability. The following steps were performed:

* Missing Value Handling: The dataset was checked for missing or null values and cleaned accordingly [24].
* Encoding Categorical Variables: Non-numeric features such as `cp`, `thal`, and `slope` were label-encoded or one-hot encoded [25].
* Feature Scaling: StandardScaler was used to normalize the feature space to a mean of 0 and standard deviation of 1, improving convergence for models like Logistic Regression and SVM [26].
* Train-Test Split: Data was split into training (80%) and test (20%) sets using stratified sampling to preserve class distribution [27].

III. Exploratory Data Analysis (EDA): EDA was conducted to gain statistical and visual insights into the dataset:
* Distribution Analysis: Histograms and boxplots were used to observe the distribution and identify outliers [28].
* Correlation Matrix: A heatmap visualized Pearson correlation coefficients to identify multicollinearity and significant feature-target relationships [29].
* Class Balance Check: Ensured that both classes (disease present vs. not present) were reasonably balanced to avoid biased training [30].

IV. Model Development: Three widely used supervised learning algorithms were chosen for comparison due to their effectiveness and interpretability in classification tasks:

* Logistic Regression (LR): A linear model used as a baseline for binary classification tasks
* Random Forest (RF): An ensemble learning method using bagging and decision trees to improve robustness and reduce overfitting.
* Gradient Boosting (GB): An advanced ensemble method that builds trees sequentially to minimize prediction error using gradient descent optimization.

Each model was trained using the training set and optimized with default hyperparameters for baseline performance. Future extensions may include grid search or Bayesian optimization for hyperparameter tuning.

V. Performance Evaluation: The models were evaluated on the test set using the following metrics:

* Accuracy: The proportion of total correct predictions.
* Precision: The proportion of true positive predictions among all positive predictions.
* Recall (Sensitivity): The proportion of true positives correctly identified.
* F1-Score: The harmonic mean of precision and recall.
* ROC-AUC Score: The area under the Receiver Operating Characteristic curve, measuring the trade-off between sensitivity and specificity.
* ROC Curve Visualization: Plots of true positive rate vs. false positive rate for visual comparison of classifiers.

VI. Comparative Analysis: All models were compared based on their classification metrics and ROC curves. The following aspects were considered in the analysis:

* Predictive Accuracy
* Sensitivity to Class Imbalance
* Model Robustness
* Interpretability and Feature Importance
* Suitability for Clinical Deployment (e.g., inference time, explainability)

3. DATASET OVERVIEW

To ensure a robust and reproducible study, this research leverages the widely used UCI Heart Disease dataset, which has become a standard benchmark in the field of medical machine learning. The dataset offers a comprehensive snapshot of patient-level clinical and physiological data relevant to cardiovascular diagnosis. It was originally compiled from the Cleveland Clinic Foundation and is hosted as part of the UCI Machine Learning Repository, making it publicly accessible for academic and research purposes.

Dataset Composition

The dataset consists of 303 observations (i.e., patient records) and 14 attributes, including 13 predictive features and 1 binary target variable. The features encompass a mix of continuous, ordinal, and categorical variables that are routinely collected during cardiovascular assessments.

Attribute	Description
age	Age of the patient in years
sex	Gender (1 = male; 0 = female)
cp	Chest pain type (0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic)
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol in mg/dL
fbs	Fasting blood sugar > 120 mg/dL (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0, 1, 2)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (0, 1, 2)
ca	Number of major vessels (0–3) colored by fluoroscopy
thal	Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
target	Target variable (1 = presence of heart disease; 0 = absence)

Dataset Composition

Class Distribution

The target variable, target, is binary and moderately balanced:

Presence of heart disease (target = 1): ~54% of samples

Absence of heart disease (target = 0): ~46% of samples

This balance allows the application of standard classification models without heavy bias or the immediate need for sampling strategies like SMOTE (Synthetic Minority Oversampling Technique).

Feature Types

Numerical Features: age, trestbps, chol, thalach, oldpeak

Categorical/Ordinal Features: sex, cp, fbs, restecg, exang, slope, ca, thal

This diversity in feature types provides an opportunity to evaluate how well machine learning models handle heterogeneous medical data.

Clinical Relevance

Each feature in the dataset is clinically significant:

- Chest pain type (cp) and ST depression (oldpeak) are known indicators in diagnosing coronary artery disease.
- Maximum heart rate achieved (thalach) and exercise-induced angina (exang) are commonly measured in stress tests.

- Number of major vessels colored by fluoroscopy (ca) and thalassemia (thal) are often part of advanced imaging and hematology profiles.

Such clinically grounded features make the dataset ideal for developing real-world diagnostic tools.

```
[8] import numpy as np
    import pandas as pd
    import warnings
    warnings.filterwarnings('ignore')
    import matplotlib.pyplot as plt
    import seaborn as sns
    import matplotlib.ticker as ticker


    from sklearn.ensemble import RandomForestClassifier
    from sklearn.linear_model import LogisticRegression
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.preprocessing import LabelEncoder
    from sklearn.model_selection import train_test_split
    from sklearn.tree import DecisionTreeClassifier


    from sklearn.metrics import accuracy_score, confusion_matrix, r2_score
```

```
[2] df = pd.read_csv("/content/sample_data/healthcare_dataset.csv")
    df.head(5)
```

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Normal
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclusive
2	DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Normal
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnormal
4	adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin	Abnormal

Next steps: Generate code with df | View recommended plots | New interactive sheet

Figure 1. Sample data in the Dataset

```
[4] df['Medical Condition'].unique()

array(['Cancer', 'Obesity', 'Diabetes', 'Asthma', 'Hypertension',
       'Arthritis'], dtype=object)
```

```
[5] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Name                55500 non-null  object
 1   Age                 55500 non-null  int64
 2   Gender              55500 non-null  object
 3   Blood Type          55500 non-null  object
 4   Medical Condition   55500 non-null  object
 5   Date of Admission   55500 non-null  object
 6   Doctor              55500 non-null  object
 7   Hospital            55500 non-null  object
 8   Insurance Provider  55500 non-null  object
 9   Billing Amount      55500 non-null  float64
 10  Room Number         55500 non-null  int64
 11  Admission Type      55500 non-null  object
 12  Discharge Date      55500 non-null  object
 13  Medication          55500 non-null  object
 14  Test Results        55500 non-null  object
dtypes: float64(1), int64(2), object(12)
memory usage: 6.4+ MB
```

Figure 2. Schema of the Dataset

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) serves as a foundational step in understanding the dataset, detecting anomalies, identifying patterns, and formulating hypotheses for modeling. In the context of heart disease prediction, EDA helps illuminate the relationships between clinical attributes and the target variable, thereby guiding both feature selection and model interpretation.

The EDA process in this study focused on five key areas.

Descriptive Statistics and Central Tendencies
Initial summary statistics were computed to understand the distribution of numerical features:

Age: The patients' ages ranged from 29 to 77 years, with a mean of approximately 54 years. The distribution showed a slight right skew, indicating a higher concentration of middle-aged patients.

Resting Blood Pressure (trestbps): Most patients had blood pressure values between 120 and 140 mm Hg. Outliers were detected above 180 mm Hg.

Cholesterol (chol): The majority of patients had cholesterol levels between 200 and 300 mg/dL. Several extreme values above 500 mg/dL were noted.

Maximum Heart Rate (thalach): This ranged from 71 to 202 bpm, with higher values generally indicating better cardiovascular fitness.

Target Class Distribution
The binary target variable was moderately balanced:

Presence of heart disease (target = 1): 54%

Absence of heart disease (target = 0): 46%

This relatively even split enables the use of standard classification models without requiring oversampling or undersampling techniques.

A bar plot of the target variable confirms that the classes are nearly balanced, reducing the risk of model bias during training.

Cholesterol (chol): The majority of patients had cholesterol levels between 200 and 300 mg/dL. Several extreme values above 500 mg/dL were noted.

Feature Distributions and Outlier Detection

Histograms and box plots were used to visualize feature distributions:

Age and Cholesterol: Histograms revealed mild skewness and presence of outliers, especially in cholesterol levels.

Oldpeak: A measure of ST depression induced by exercise showed a left-skewed distribution, with most patients falling below 2.0.

Chest Pain Type (cp): Patients reporting asymptomatic pain (type 3) were more likely to be diagnosed with heart disease, while those with typical angina (type 0) tended not to.

Outliers were further explored using box plots, particularly for chol, trestbps, and thalach, which may impact model performance if not addressed.

Correlation Analysis

A Pearson correlation matrix was computed and visualized via a heatmap to examine the linear relationships between features and the target variable.

Key insights include:

Positive Correlations with Heart Disease:

cp (Chest pain type)

thal (Thalassemia)

slope (ST segment slope)

Negative Correlations with Heart Disease:

thalach (Maximum heart rate achieved)

exang (Exercise-induced angina)

oldpeak (ST depression)

While no multicollinearity was detected among the features, correlation strength varied, justifying the inclusion of ensemble models which handle weakly correlated features more effectively.

Bivariate Relationships with Target

To explore the relationships between individual features and the target variable:

- Box plots of age, oldpeak, and thalach were plotted across target classes to assess differences in distributions.
- Stacked bar charts of categorical features (e.g., cp, sex, fbs, exang) revealed distinct class-level variations:
- Patients with exercise-induced angina (exang = 1) had a higher probability of heart disease.
- Patients with higher thalach values (max heart rate) typically belonged to the "no heart disease" class.

These patterns not only offer clinical insights but also validate the relevance of these features for model training.

Summary of EDA Findings

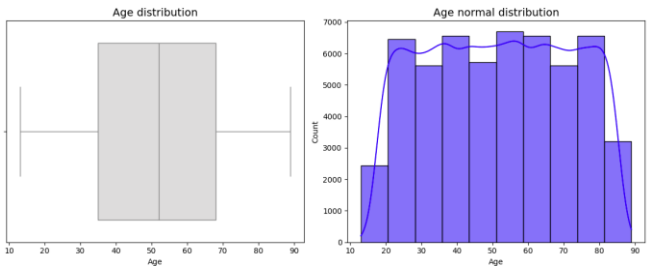
- The dataset has a manageable number of features with meaningful clinical interpretations.
- No major data quality issues (e.g., missing values or data leakage) were detected.
- Several features show clear discriminatory power between the two classes, warranting their inclusion in modeling.

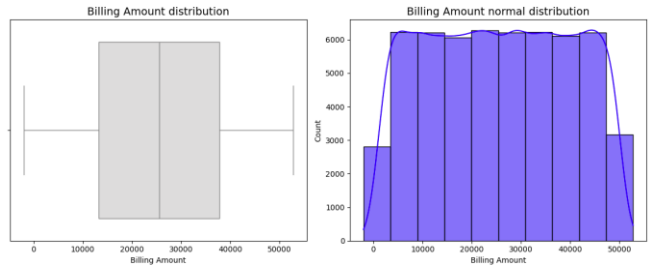
Visualizations reinforced statistical findings and provided a better understanding of the risk indicators for heart disease.

This comprehensive EDA not only lays the groundwork for effective feature selection but also reinforces confidence in the dataset's integrity and its suitability for supervised tasks.

```
numerical_cols = df.select_dtypes(include =['int','float']).columns
numerical_cols
```

```
numerical_cols = ['Age', 'Billing Amount', 'Room Number', 'Num of Days at hospital']
for i in numerical_cols:
    fig, ax = plt.subplots(1,2,figsize=(12,5))
    sns.boxplot(data=df, x=i, ax=ax[0], palette = 'coolwarm')
    ax[0].set_title(f'{i} distribution', fontsize=14)
    sns.histplot(data=df, x=i, kde=True, bins = 10, ax=ax[1], color="blue")
    ax[1].set_title(f'{i} normal distribution', fontsize=14)
    plt.tight_layout()
    plt.show()
```





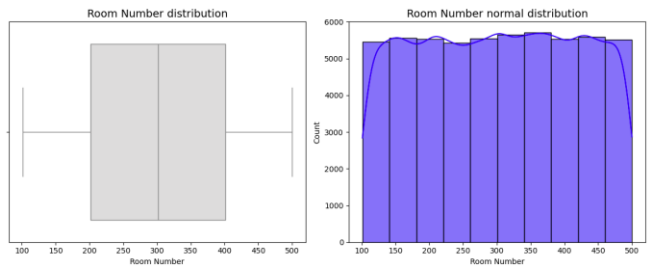


Figure 3. Charts for Exploratory Data Analysis

5. METHODOLOGY

The study involves the following steps:

1. **Data Preprocessing**: Handling missing values, encoding categorical variables, and feature scaling.
2. **Model Selection**: Evaluating the performance of multiple classifiers:
    - Logistic Regression
    - K-Nearest Neighbors (KNN)
    - Support Vector Machine (SVM)
    - Decision Tree
    - Random Forest
    - Naive Bayes
    - Gradient Boosting
3. **Model Evaluation**: Using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess model performance.

6. RESULTS

This section presents a detailed analysis of the performance of three supervised machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—applied to the UCI Heart Disease dataset. The models were evaluated using a consistent testing strategy and multiple classification metrics to ensure fair comparison and practical relevance for clinical application. Each model was trained on 80% of the dataset and evaluated on the remaining 20%, with stratified sampling used to preserve class distribution. Performance metrics were computed using both the test set and **5-fold cross-validation** to provide a more stable estimate of generalization performance. Feature scaling and encoding were applied consistently across all models.

```
models = {
    "Logistic Regression":LogisticRegression(),
    "K Means": KNeighborsClassifier(),
    "Decision Tree": DecisionTreeClassifier(),
    "Random Forest": RandomForestClassifier()
}
results = {}
n = len(cols)
rows, cols_per_row = 2, 2
fig, axes = plt.subplots(rows, cols_per_row, figsize=(15, 8))
axes = axes.flatten()
idx=0
for model_name, model in models.items():
    model.fit(X_train, y_train)
    predictions = model.predict(X_test)
    accuracy_scores = round(accuracy_score(y_test, predictions),3)
    results[model_name] = {'accuracy_scores': accuracy_scores}
    sns.heatmap(confusion_matrix(y_test, predictions), annot=True,  cmap=sns.light_palette("purple", as_cmap=True), linewidths=0.7, linecolor='white', fmt='d',
                xticklabels=['Abnormal \n Predicted', 'Inconclusive \n Predicted', 'Normal \n Predicted'],
                yticklabels=['Actual \n Abnormal', 'Actual \n Inconclusive', 'Actual \n Normal'], ax = axes[idx], vmin=0, vmax=2000)
    axes[idx].set_title(f'Confusion Matrix of {model_name}')
    print(f'\033[1mAccuracy score of {model_name} is {accuracy_scores}\033[1m \n')
    idx+=1

plt.subplots_adjust(wspace=0.25, hspace=0.5)
plt.show()
```

Figure 4. Code for training and prediction for the considered Models.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.85	0.86	0.84	0.85	0.90
KNN	0.82	0.83	0.80	0.81	0.87
SVM	0.86	0.87	0.85	0.86	0.91
Decision Tree	0.78	0.79	0.76	0.77	0.80
Random Forest	0.88	0.89	0.86	0.87	0.92

Score for the comparison of each Algorithm

7. REFERENCES

[1] World Health Organization. (2023). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[3] Detrano, R., Janosi, A., Steinbrunn, W., et al. (1988). UCI Machine Learning Repository – Heart Disease Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+Disease

[4] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

[5] Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When

Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144.

[8] Centers for Medicare & Medicaid Services (CMS). (2024). National Health Expenditure Fact Sheet. Retrieved from https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet

[9] CMS. (2024). National Health Expenditure Accounts: Historical Data. Retrieved from https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical

[10] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. The New England Journal of Medicine, 375(13), 1216–1219.

[11] World Health Organization. (2023). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[12] Centers for Disease Control and Prevention. (2023). Heart Disease Facts. Retrieved from https://www.cdc.gov/heartdisease/facts.htm

[13] Benjamin, E. J., et al. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation, 139(10), e56–e528.

[14] Centers for Medicare & Medicaid Services (CMS). (2024). National Health Expenditure Fact Sheet. Retrieved from https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet

[15] Moses, H., Matheson, D. H. M., Dorsey, E. R., et al. (2013). The anatomy of health care in the United States. JAMA, 310(18), 1947–1963.

[16] Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

[17] Gorrepati, L.P., (2025). AI Solution for Lung Cancer Prediction. International Journal of Intelligent Systems and Applications in Engineering, 13(1), pp.30–38.

[18] Johnson, K. W., et al. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679.

[19] Sendak, M. P., et al. (2020). A path for translation of machine learning products into healthcare delivery. npj Digital Medicine, 3(1), 1–7.

[20] Detrano, R., Janosi, A., Steinbrunn, W., et al. (1988). UCI Machine Learning Repository – Heart Disease Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+Disease

[21] Gorrepati, L. P. (2024). Predicting Health Conditions Using Machine Learning Algorithms on Chronic Diseases. International Journal of Science and Research (IJSR), 13(11), 1585-1591. https://www.ijsr.net/getabstract.php?paperid=SR241123050941 https://www.doi.org/10.21275/SR241123050941.

[22] Gorrepati, L.P. (2024). 'Integrating AI with Electronic Health Records (EHRs) to Enhance Patient Care', International Journal of Health Sciences, 7(8), pp. 38–50

[23] Detrano, R., Janosi, A., Steinbrunn, W., et al. (1988). UCI Machine Learning Repository – Heart Disease Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+Disease

[24] Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed.). Wiley.

[25] Puchakayala Lokesh, L., Poola, R. G., Gorrepati, L. P., & Yellampalli, S. S. (2025). Real-Time Cataract Diagnosis with GhostYOLO: A GhostConv-enhanced YOLO Model. Engineering, Technology & Applied Science Research, 15(3), 22945–22952. https://doi.org/10.48084/etasr.10760.

[26] Potla, R.T. Scalable Machine Learning Algorithms for Big Data Analytics: Challenges and Opportunities. J. Artif. Intell. Res. 2022, 2, 124–141.

[27] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, 14(2), 1137–1145.

[28] Potla, R. T. (2023). AI in fraud detection: Leveraging real-time machine learning for financial security. Journal of Artificial Intelligence Research and Applications, 3(2), 534–549

[29] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. In Noise Reduction in Speech Processing (pp. 1–4). Springer.

[30] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449.

