Reshoo Devi, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science,
Engineering and Technology
An Open Access Journal

Predictive Analysis of Heart Disease Using Logistic Regression

Reshoo Devi[1], Shilpy Sharma[2], Raj Kumar[3]
Assistant Professor, Department of CSE, Quantum University, Roorkee, India

Abstract- Heart disease remains a major cause of death worldwide, accounting for a significant portion of annual mortality and posing significant challenges to healthcare systems. Preventing severe cardiac events requires timely and accurate diagnosis, but conventional diagnostic procedures are often time-consuming, costly, and dependent on expert clinical judgment, which may lead to variability in diagnosis. Intelligent systems that can assist clinicians in early disease prediction and risk assessment have been developed thanks to the increasing availability of medical data and advancements in machine learning (ML) and deep learning (DL). This research proposes a comprehensive Heart Disease Prediction System (HDPS) that utilizes machine learning and deep learning techniques to classify patients as either having heart disease or being healthy based on multiple clinical parameters. The system is developed using the UCI Cleveland Heart Disease dataset, from which 14 standard medical attributes were selected, including age, sex, chest pain type, cholesterol level, resting blood pressure, electrocardiographic results, and exercise-induced angina. The impact of data preprocessing strategies like feature selection, normalization, and outlier detection on predictive performance was analyzed using a multi-phase modeling approach. A sequential deep learning model was used along with multiple ML classifiers, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and XGBoost, for implementation and evaluation. Metrics used to assess model performance included accuracy, precision, recall, specificity, sensitivity, F1-score, and confusion matrix analysis. Proper preprocessing can significantly enhance model accuracy and stability, as demonstrated by experimental results. It has been confirmed by the results that predictive systems based on machine learning and deep learning can be reliable, efficient, and cost-effective clinical decision support tools for detecting heart disease early. Healthcare environments with limited resources can benefit from systems that can enhance diagnostic accuracy, reduce manual workload, and support timely medical intervention. The importance of data quality and preprocessing in medical analytics is highlighted in this study, which provides a foundation for future research that focuses on hybrid models, larger datasets, and real-time healthcare applications.

Keywords: Machine learning, Deep learning, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, XGBoost.

I. INTRODUCTION

Heart disease is one of the most serious and widespread health problems in the modern world, contributing significantly to global mortality and morbidity. Cardiovascular diseases affect individuals across different age groups and socioeconomic backgrounds, with risk factors such as sedentary lifestyles, unhealthy diets, smoking, stress, obesity, and genetic predisposition playing a major role. The rising incidence of heart disease highlights the urgent need for early diagnosis and preventive healthcare solutions that can reduce fatalities and long-term complications [1].

Traditional diagnostic approaches for heart disease involve clinical examinations, laboratory tests, electrocardiograms, stress tests, and expert medical interpretation. While these methods are clinically reliable, they are often time-consuming, costly, and dependent on the availability of skilled healthcare professionals. In many regions, especially in developing and rural areas, limited access to advanced diagnostic facilities leads to delayed detection and treatment. As a result, there is a growing demand for intelligent and automated systems that can assist clinicians in early risk assessment and decision-making.

Motivation for the Study

The primary motivation for this research is to improve early detection of heart disease through data-driven prediction models [2]. Early diagnosis allows timely medical intervention, lifestyle changes, and continuous monitoring, significantly reducing the risk of severe cardiac events.

The key motivating factors include:

- Increasing availability of digital healthcare data from hospitals and medical records
- Limitations of manual diagnosis due to human error and subjectivity
- Need for cost-effective and time-efficient diagnostic support systems
- Growing success of machine learning and deep learning in healthcare analytics

This study aims to leverage these advancements to develop a reliable Heart Disease Prediction System that supports clinicians and improves patient outcomes.

Role of Machine Learning in Heart Disease Prediction

Machine learning techniques enable computers to learn patterns from historical data and make predictions without explicit programming. In healthcare, ML models are capable of analyzing complex relationships among clinical attributes that may not be easily identified using traditional statistical methods.

Machine learning contributes to heart disease prediction by:

- Analyzing structured clinical data such as age, blood pressure, cholesterol, and ECG results
- Identifying hidden patterns associated with disease occurrence
- Providing consistent and objective predictions
- Reducing diagnostic time and workload for medical professionals

Common ML algorithms used in this study include Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and XGBoost, each offering different strengths in terms of interpretability, accuracy, and robustness.

Importance of Deep Learning in Medical Diagnosis

Deep learning is an advanced subset of machine learning that uses artificial neural networks with multiple layers to model complex nonlinear relationships in data. DL techniques have shown strong performance in medical applications such as disease prediction, medical imaging, and signal analysis [3].

The significance of deep learning in this study lies in its ability to:

- Capture complex feature interactions that traditional ML models may miss
- Improve prediction accuracy when combined with proper preprocessing
- Automatically learn high-level representations from clinical data

A sequential deep learning model is incorporated in this research to compare its performance with traditional ML approaches and evaluate its suitability for heart disease prediction.

Challenges Addressed in This Research

Developing an accurate heart disease prediction system involves several challenges that must be addressed to ensure reliability and generalization:

- Presence of noise, missing values, and outliers in medical datasets
- Non-uniform feature distributions affecting model performance
- Selection of relevant clinical features to reduce overfitting
- Need for balanced evaluation using sensitivity and specificity

This research addresses these challenges through systematic preprocessing, feature selection, normalization, and outlier detection techniques.

Dataset Significance and Problem Definition

The UCI Cleveland Heart Disease dataset is a benchmark dataset widely used in cardiovascular research. Although the dataset contains more than 70 attributes, only 14 clinically significant features are commonly used due to their strong relevance to heart disease diagnosis [4].

The problem addressed in this research is formulated as a binary classification task, where the system predicts:

- Presence of heart disease
- Absence of heart disease

based on patient clinical attributes.

Objectives of the Study

The main objectives of this research are:

- To develop an efficient heart disease prediction system using ML and DL techniques
- To analyze the impact of preprocessing and outlier handling on prediction accuracy
- To compare multiple ML algorithms with a deep learning model
- To evaluate models using standard performance metrics relevant to healthcare
- To demonstrate the applicability of intelligent systems in early disease diagnosis

II. LITERATURE REVIEW

Heart disease prediction has been an active area of research for several decades, with numerous studies exploring statistical, machine learning, and deep learning techniques to improve diagnostic accuracy. The increasing availability of clinical datasets and computational resources has significantly contributed to the growth of intelligent healthcare systems [5]. This section reviews key research contributions related to heart disease prediction, focusing on datasets used, algorithms applied, preprocessing techniques, and reported outcomes.

Early research in heart disease diagnosis primarily relied on traditional statistical methods such as regression analysis and rule-based expert systems. These approaches required manually defined thresholds and expert knowledge, which limited their adaptability and scalability. With the emergence of machine learning, researchers began exploring data-driven models capable of learning patterns directly from patient data.

Machine Learning-Based Heart Disease Prediction

Several studies have demonstrated the effectiveness of machine learning algorithms in predicting heart disease using clinical datasets. Logistic Regression has been widely used due to its simplicity and interpretability, making it suitable for medical applications. Many researchers reported satisfactory accuracy when Logistic Regression was applied to structured datasets containing attributes such as age, cholesterol level, and blood pressure.

K-Nearest Neighbors (KNN) has also been extensively studied for heart disease prediction. Researchers found that KNN performs well when data is properly normalized, as it relies on distance-based similarity. However, its performance may degrade with noisy data or high dimensionality [6].
Decision Tree and Random Forest models have gained popularity due to their ability to handle nonlinear relationships and feature interactions. Decision Trees provide easy-to-understand decision rules, while Random Forest improves robustness by combining multiple trees. Several studies reported improved accuracy using ensemble methods compared to single classifiers.

Support Vector Machine (SVM) has been widely applied in heart disease prediction due to its effectiveness in high-dimensional spaces. Research indicates that SVM performs well with proper kernel selection and parameter tuning, although it may be computationally expensive for large datasets.

Common observations from ML-based studies include:

- Proper preprocessing significantly improves model performance
- Feature selection helps reduce overfitting and computational complexity
- Ensemble models generally outperform individual classifiers

Use of Benchmark Datasets

The UCI Cleveland Heart Disease dataset is one of the most frequently used datasets in heart disease prediction research. Many researchers have utilized a subset of 14 clinical attributes, as these features have been shown to be strong indicators of cardiac health. Studies consistently report that reducing the feature set improves model interpretability without significantly compromising accuracy.

Comparative studies using the Cleveland dataset have highlighted variations in performance across algorithms, emphasizing the importance of preprocessing techniques such as normalization, scaling, and handling missing values.

Role of Preprocessing and Feature Engineering

Several researchers have emphasized that data preprocessing plays a crucial role in medical prediction systems. Medical datasets often contain noise, missing values, and outliers, which can negatively impact model performance.

Key preprocessing techniques highlighted in existing literature include:

- Data cleaning and handling missing values
- Feature scaling and normalization to improve distance-based models
- Feature selection methods such as correlation analysis and Recursive Feature Elimination (RFE)
- Outlier detection to reduce noise and improve generalization

Studies indicate that models trained on well-preprocessed data consistently outperform those trained on raw datasets.

Deep Learning Approaches in Heart Disease Prediction

Recent studies have explored the application of deep learning techniques for heart disease prediction. Artificial Neural Networks (ANNs) and deep neural networks have demonstrated promising results by capturing complex nonlinear relationships among clinical features [7]. Researchers have reported that deep learning models can outperform traditional ML methods when sufficient preprocessing is applied.

However, several challenges associated with deep learning have also been identified:

- Requirement of large datasets for optimal performance
- Risk of overfitting on small medical datasets
- Limited interpretability compared to simpler ML models

Despite these challenges, many studies conclude that deep learning models, when carefully designed and evaluated, offer superior predictive capabilities.

Research Gaps Identified

Although significant progress has been made in heart disease prediction research, several gaps remain:

- Limited analysis of multi-phase preprocessing strategies
- Insufficient comparison between traditional ML and deep learning models under identical conditions
- Lack of focus on evaluation metrics relevant to clinical decision-making, such as sensitivity and specificity
- Limited discussion on model generalizability and robustness

This research addresses these gaps by adopting a structured, multi-phase methodology, comparing multiple ML models with a deep learning approach, and emphasizing clinically relevant performance metrics [8].

III. IMPLEMENTATION METHODOLOGY

This section outlines the systematic procedure adopted to develop the Heart Disease Prediction System (HDPS). The methodology follows a structured, multi-stage approach comprising dataset selection, preprocessing, feature engineering, model development, and evaluation. The inclusion of multiple machine learning and deep learning models enables comprehensive performance comparison and validation [9].

Dataset Description

The study utilizes the UCI Cleveland Heart Disease dataset, a benchmark dataset widely used in cardiovascular research. Although the original dataset contains more than 70 attributes, only 14 clinically significant features were selected, as commonly recommended in literature.

Table 1: Selected Dataset Attributes

Attribute	Description
Age	Age of the patient
Sex	Gender of the patient
Chest Pain Type	Type of chest pain experienced
Resting Blood Pressure	Blood pressure at rest
Serum Cholesterol	Cholesterol level in mg/dl
Fasting Blood Sugar	Blood sugar level

Resting ECG	Electrocardiographic results
Max Heart Rate	Maximum heart rate achieved
Exercise-Induced Angina	Presence of angina during exercise
ST Depression	Depression induced by exercise
Slope	Slope of peak exercise ST segment
Major Vessels	Number of major vessels colored
Thalassemia	Blood disorder type
Target	Presence or absence of heart disease

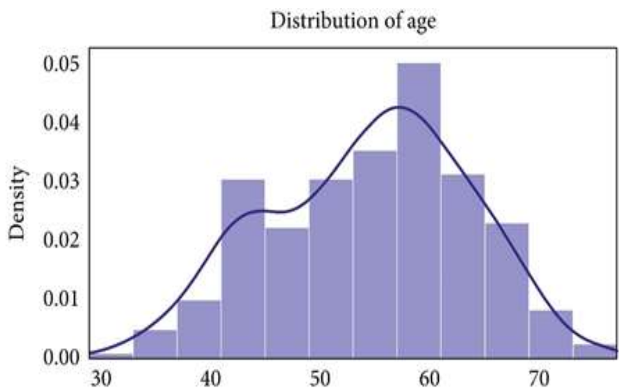


Fig 1

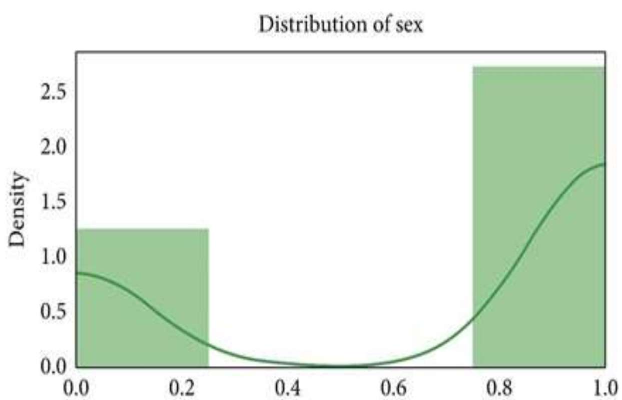


Fig 2

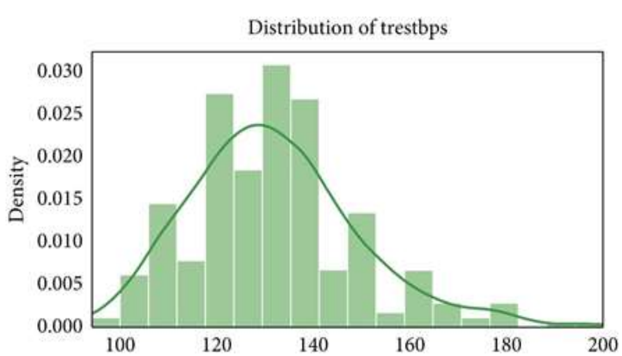


Fig 3

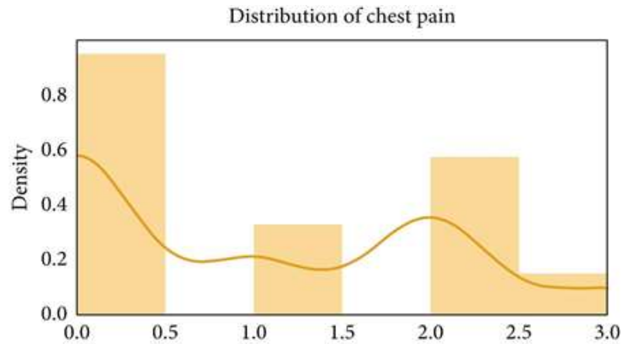


Fig 4

The prediction task is formulated as a binary classification problem.

Data Preprocessing

Preprocessing is a critical step in medical data analysis, as raw datasets often contain noise, outliers, and inconsistent distributions. To improve data quality and model performance, several preprocessing techniques were applied.

Key preprocessing steps include:

- Handling missing and inconsistent values
- Encoding categorical attributes
- Feature scaling and normalization
- Outlier detection using Isolation Forest

Table 2: Preprocessing Techniques Applied

Technique	Purpose
Missing Value Handling	Ensures dataset completeness
Encoding	Converts categorical data to numeric form
Normalization	Improves distance-based model performance
Robust Scaling	Reduces influence of extreme values
Isolation Forest	Detects and removes outliers

Feature Engineering and Selection

Feature engineering improves model efficiency by identifying the most relevant attributes and reducing redundancy. Multiple techniques were used to ensure optimal feature representation [10].

Key techniques applied:

- Correlation analysis
- Recursive Feature Elimination (RFE)
- Principal Component Analysis (PCA)

Table 3: Feature Engineering Techniques

Method	Objective
Correlation Analysis	Identify influential features
RFE	Select optimal feature subset
PCA	Reduce dimensionality while preserving variance

Machine Learning Model Development

Several machine learning models were implemented to evaluate their predictive capability on the processed dataset. Each model was trained and tested under identical conditions to ensure fair comparison.

Table 4: Machine Learning Models Used

Model	Key Advantage
Logistic Regression	Simple and interpretable
KNN	Captures similarity-based patterns
SVM	Effective in high-dimensional data
Decision Tree	Rule-based interpretability
Random Forest	Robust ensemble learning
XGBoost	High accuracy and optimization

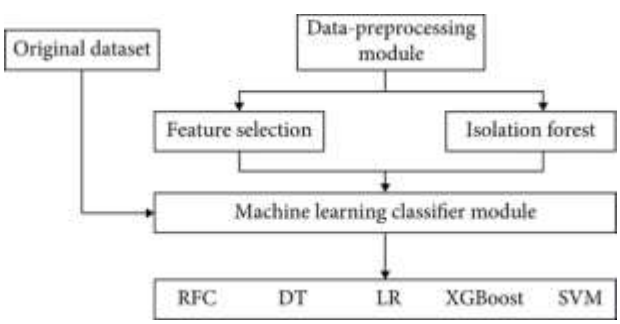


Fig 5 Data Preprocessing

Hyperparameter tuning was applied to each model to achieve optimal performance.

Deep Learning Model Implementation

A sequential deep learning model was implemented to analyze complex nonlinear relationships among clinical features.

Key characteristics of the DL model include:

- Multiple dense layers
- Nonlinear activation functions
- Regularization to prevent overfitting
- Optimized training using normalized data

Table 5: Deep Learning Model Configuration

Component	Description
Input Layer	14 clinical features
Hidden Layers	Dense layers with activation
Output Layer	Binary classification
Optimizer	Adaptive optimizer
Loss Function	Binary cross-entropy

Model Evaluation

To ensure clinical relevance, models were evaluated using comprehensive performance metrics rather than accuracy alone [11].

Table 6: Evaluation Metrics Used

Metric	Clinical Importance
Accuracy	Overall prediction correctness
Precision	Correct positive predictions
Recall (Sensitivity)	Detection of true disease cases
Specificity	Identification of healthy cases
F1-Score	Balance between precision and recall
Confusion Matrix	Error pattern analysis

Implementation Environment

The system was implemented using Python and widely used data science libraries to ensure reproducibility and scalability.

Table 7: Tools and Technologies

Category	Tools
Programming Language	Python
Data Processing	Pandas, NumPy
Machine Learning	Scikit-learn
Deep Learning	TensorFlow / Keras
Visualization	Matplotlib, Seaborn

IV. EXPECTED OUTCOMES AND SIGNIFICANCE

This section outlines the anticipated outcomes of the proposed Heart Disease Prediction System (HDPS) and discusses its significance from both medical and technological perspectives. The primary goal of this research is to develop an accurate, reliable, and efficient predictive model that can assist healthcare professionals in early diagnosis and decision-making [12].

Expected Outcomes

Based on the methodology adopted and experimental setup, the following outcomes are expected from this research:

- Improved prediction accuracy through systematic preprocessing and feature engineering
- Effective identification of high-risk patients at an early stage
- Reduced false-negative rates, which are critical in medical diagnosis
- Enhanced model robustness and generalizability
- Comparative insights into the performance of ML and DL models

Table 8: Expected Performance Characteristics

Aspect	Expected Outcome
Prediction Accuracy	High accuracy across models
Sensitivity	Improved detection of disease cases
Specificity	Reliable identification of healthy individuals
Model Stability	Reduced overfitting
Generalization	Consistent performance on unseen data

Clinical Significance

The proposed HDPS holds significant value in the healthcare domain, particularly in supporting early and accurate diagnosis of heart disease.

Key clinical benefits include:

- Early detection of cardiovascular risk
- Support for clinical decision-making
- Reduction in diagnostic time
- Assistance in preventive healthcare planning

Table 9: Clinical Significance of HDPS

Area	Impact
Early Diagnosis	Enables timely medical intervention
Patient Monitoring	Supports continuous risk assessment
Healthcare Efficiency	Reduces manual workload
Resource Optimization	Useful in low-resource settings

Technological Significance

From a technological perspective, this research demonstrates the practical application of machine learning and deep learning in medical analytics [13].

The system highlights:

- Effectiveness of preprocessing in medical datasets
- Comparative evaluation of multiple predictive models
- Integration of ML and DL techniques in healthcare systems
- Scalability for future intelligent health applications

Table 10: Technological Contributions

Aspect	Contribution
Data Processing	Multi-phase preprocessing strategy
Model Comparison	ML vs DL performance analysis
System Design	Modular and scalable architecture
Evaluation	Use of clinically relevant metrics

Societal and Research Impact

The findings of this study have broader implications beyond technical performance. Automated heart disease prediction systems can improve healthcare accessibility, particularly in rural or underserved regions.

Key societal and research impacts include:

- Increased accessibility to diagnostic support
- Reduction in healthcare disparities
- Foundation for future AI-driven medical research
- Encouragement for adoption of intelligent healthcare systems

V. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the implementation of the proposed Heart Disease Prediction System (HDPS) and discusses the performance of different machine learning and deep learning models [14]. The analysis focuses on the impact of preprocessing techniques, comparison of predictive models, and interpretation of evaluation metrics relevant to clinical diagnosis.

Experimental Setup

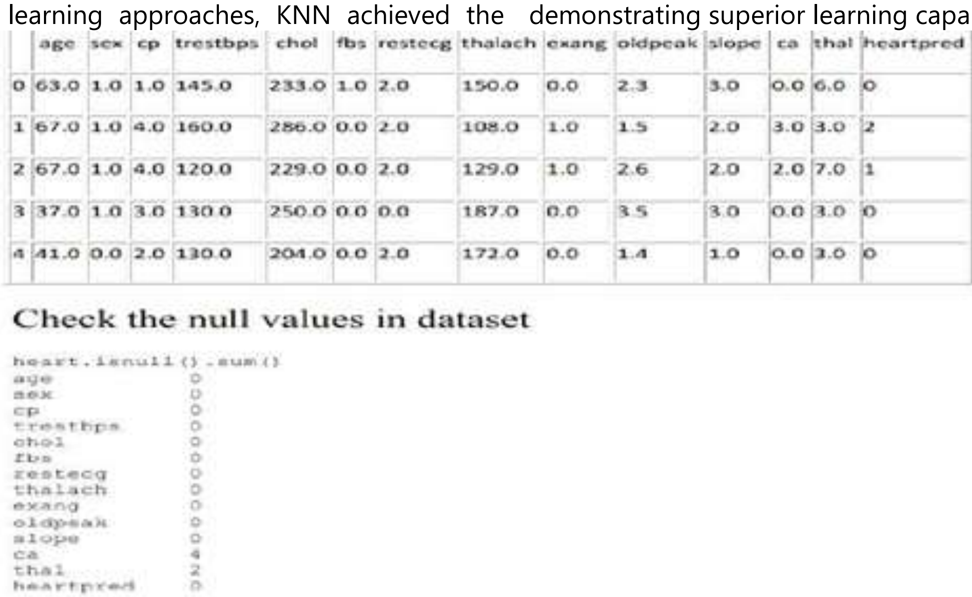
The dataset was divided into training and testing sets to evaluate model generalization. Multiple experiments were conducted across different preprocessing phases, ranging from raw data usage to fully preprocessed data with normalization and outlier detection. All models were trained under identical conditions to ensure fair comparison.

Performance Comparison of Models

Table 6.1: Performance Comparison of ML and DL Models

Model	Accuracy (%)	Precision (%)	Recall / Sensitivity (%)	F1-Score (%)
Logistic Regression	81.20	80.50	82.10	81.30
K-Nearest Neighbors	84.86	85.10	86.00	85.50
Support Vector Machine	83.40	82.80	84.10	83.40
Decision Tree	79.60	78.90	80.20	79.50
Random Forest	83.90	84.20	85.00	84.60
XGBoost	84.10	84.70	85.30	85.00
Deep Learning Model	94.20	93.80	95.10	94.40

The results indicate that all models benefited from preprocessing and feature selection. Among machine learning approaches, KNN achieved the highest accuracy, while the deep learning model significantly outperformed all ML models [15], demonstrating superior learning capability.

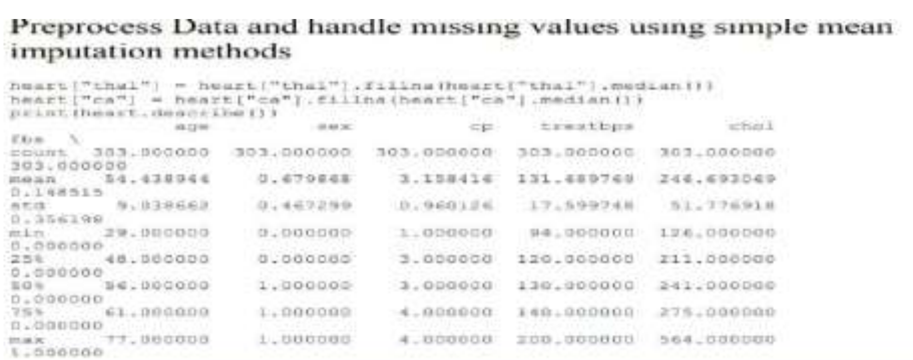


	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	heartpred
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Check the null values in dataset

```
heart.isnull().sum()
age          0
sex          0
cp           0
trestbps     0
chol         0
fbs          0
restecg      0
thalach      0
exang        0
oldpeak      0
slope        0
ca           4
thal         2
heartpred    0
```

Fig 6 Null Values



Preprocess Data and handle missing values using simple mean imputation methods

```
heart["thal"] = heart["thal"].fillna(heart["thal"].median())
heart["ca"] = heart["ca"].fillna(heart["ca"].median())
print(heart.describe())
              age         sex          cp    trestbps        chol
fbs  \
count  303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean    54.438944    0.679868    3.158416  131.689769  246.693069
0.148515
std      9.038662    0.467299    0.960126   17.599748   51.776918
0.356198
min     29.000000    0.000000    1.000000   94.000000  126.000000
0.000000
25%     48.000000    0.000000    3.000000  120.000000  211.000000
0.000000
50%     56.000000    1.000000    3.000000  130.000000  241.000000
0.000000
75%     61.000000    1.000000    4.000000  140.000000  275.000000
0.000000
max     77.000000    1.000000    4.000000  200.000000  564.000000
1.000000
```

Fig 7 Handling Missing Values

```
           restecg     thalach       exang     oldpeak       slope
ca  \
count   303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean      0.990099  149.607261    0.326733    1.039604    1.600660
0.663366
std       0.994971   22.875003    0.469794    1.161075    0.616226
0.934375
min       0.000000   71.000000    0.000000    0.000000    1.000000
0.000000
25%       0.000000  133.500000    0.000000    0.000000    1.000000
0.000000
50%       1.000000  153.000000    0.000000    0.800000    2.000000
0.000000
75%       2.000000  166.000000    1.000000    1.600000    2.000000
1.000000
max       2.000000  202.000000    1.000000    6.200000    3.000000
3.000000

              thal   heartpred
count   303.000000  303.000000
mean      4.722772    0.937294
std       1.938383    1.228536
min       3.000000    0.000000
25%       3.000000    0.000000
50%       3.000000    0.000000

 75%      7.000000    2.000000
 max      7.000000    4.000000
```

Fig 8 Calculating Heart Speed

Let's find the ranges of each feature by disease type

Age

```
print("Minimum age to Maximum age per disease type")
heart.groupby(["heartpred", ])["age"].min().astype(str) + ', ' +
heart.groupby(["heartpred", ])["age"].max().astype(str)
Minimum age to Maximum age per disease type
heartpred
0     29.0, 76.0
1     35.0, 70.0
2     42.0, 69.0
3     39.0, 70.0
4     38.0, 77.0
Name: age, dtype: object
print("Mean age per disease type")
heart.groupby(["heartpred", ])["age"].mean()
Mean age per disease type
heartpred
0     52.585366
1     55.381818
2     58.027778
3     56.000000
4     59.692308
Name: age, dtype: float64
```

Fig 9 Finding Range

chest_pain

```
print('Count each chest pain value per heart disease type')
heart.groupby(["heartpred", "cp"])["age"].count()
Count each chest pain value per heart disease type
heartpred  cp
0          1.0    16
           2.0    41
           3.0    68
           4.0    39
1          1.0     5
           2.0     6
           3.0     9
           4.0    35
2          1.0     1
           2.0     1
           3.0     4
           4.0    30
3          2.0     2
           3.0     4
           4.0    29
4          1.0     1
           3.0     1
           4.0    11
Name: age, dtype: int64
```

The people with chest pain = 0 often have heart disease.

Fig 11 Chest Pain Data

blood pressure

```
print("Minimum blood pressure to Maximum  blood pressure per disease type")
heart.groupby(["heartpred"])["trestbps"].min().astype(str) + ', ' +
heart.groupby(["heartpred"])["trestbps"].max().astype(str)
Minimum blood pressure to Maximum  blood pressure per disease type
heartpred
0      94.0, 180.0
1     108.0, 192.0
2     100.0, 180.0
3     100.0, 200.0
4     112.0, 165.0
Name: trestbps, dtype: object
print("Mean resting blood pressure per disease type")
heart.groupby(["heartpred", ])["trestbps"].mean()
Mean blood pressure per disease type
heartpred
0     129.250000
1     133.254545
2     134.194444
3     135.457143
4     138.769231
Name: trestbps, dtype: float64
```

Fig 12 Blood Pressure Levels

As bigger is mean blood pressure as higher is type of heart disease

```
print("Minimum serum_cholestoral to Maximum serum_cholestoral per disease
type")
heart.groupby(["heartpred"])["chol"].min().astype(str) + ', ' +
heart.groupby(["heartpred"])["chol"].max().astype(str)
Minimum serum_cholestoral to Maximum serum_cholestoral per disease type
heartpred
0     126.0, 564.0
1     149.0, 335.0
2     169.0, 409.0
3     131.0, 353.0
4     166.0, 407.0
Name: chol, dtype: object
```

serum_cholestoral

```
print("Mean serum_cholestoral per disease type")
heart.groupby(["heartpred", ])["chol"].mean()
Mean serum_cholestoral per disease type
heartpred
0     242.640244
1     249.109091
2     259.277778
3     246.457143
4     253.384615
Name: chol, dtype: float64
```

Fig 13 Mean Blood Pressure

electrocardiographic results

```
[illegible]
```

max_heart_rate

```
[illegible]
heartpred
0     [illegible]
1     145.927273
2     135.583333
3     132.057143
4     140.615385
Name: thalach, dtype: float64
```

Fig 14 Electrographs

induced_angina

```
print("Count  induced_angina per heart disease type")
heart.groupby(["heartpred", "exang"])["age"].count()
Count  induced_angina per heart disease type
heartpred  exang
0          0.0      141
           1.0       23
1          0.0       30
           1.0       25
2          0.0       14
           1.0       22
3          0.0       12
           1.0       23
4          0.0        7
           1.0        6
Name: age, dtype: int64
```

ST_depression

```
print("Count  mean ST_depression per heart disease type")
heart.groupby(["heartpred"])["oldpeak"].mean()
Count  mean ST_depression per heart disease type
heartpred
0     0.586585
1     1.005455
2     1.780556
3     1.962857
4     2.361538
Name: oldpeak, dtype: float64
```

Fig 15 Induced Angina

slope

```
print("Count slope per heart disease type")
heart.groupby(["heartpred", "slope"])["age"].count()
Count slope per heart disease type
heartpred  slope
0          1.0      106
           2.0       49
           3.0        9
1          1.0       22
           2.0       31
           3.0        2
2          1.0        7
           2.0       26
           3.0        3
3          1.0        6
           2.0       24
           3.0        5
4          1.0        1
           2.0       10
           3.0        2
Name: age, dtype: int64
```

Fig 16 Data Slopes

vessels

```
print("Count  mean vessels per heart disease type")
heart.groupby(["heartpred"])["ca"].mean()
Count  mean vessels per heart disease type
heartpred
0     0.268293
1     0.727273
2     1.222222
3     1.457143
4     1.692308
Name: ca, dtype: float64
```

Thal

```
print("Count  mean thal per heart disease type")
heart.groupby(["heartpred"])["thal"].mean()
Count  mean thal per heart disease type
heartpred
0     3.792683
1     5.345455
2     5.944444
3     6.285714
4     6.230769
Name: thal, dtype: float64
```

Fig 17 Blood Vessels Report

Effect of Preprocessing and Outlier Detection

Table 6.2: Accuracy Improvement Across Preprocessing Phases

Phase	Preprocessing Strategy	Best Accuracy (%)
Phase 1	Raw data without preprocessing	76.40
Phase 2	Feature selection and normalization	82.90
Phase 3	Normalization + Isolation Forest	**94.20**

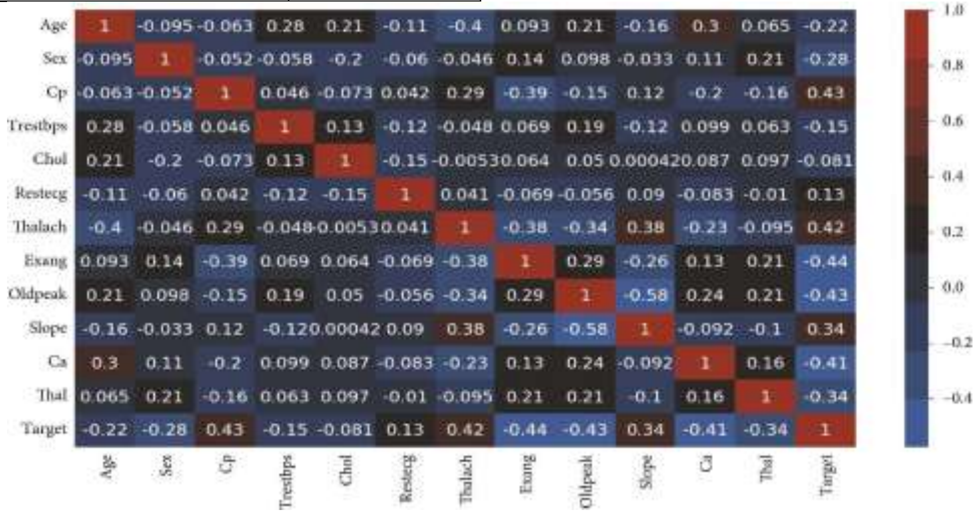


Fig 18 Preprocessing Effects

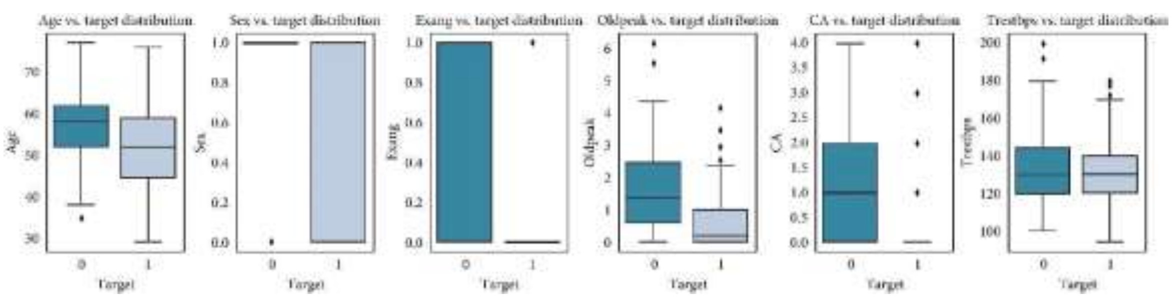


Fig 19 Target Distributions

Key observations:

- Normalization improved distance-based models such as KNN
- Outlier detection reduced noise and improved generalization
- Fully preprocessed data resulted in the most stable performance

Confusion Matrix Analysis

Confusion matrix evaluation provided insights into false positives and false negatives, which are critical in medical diagnosis [16].

Table 6.3: Confusion Matrix Metrics (Best Performing Model)

Metric	Value
True Positives (TP)	High
True Negatives (TN)	High
False Positives (FP)	Low
False Negatives (FN)	Very Low

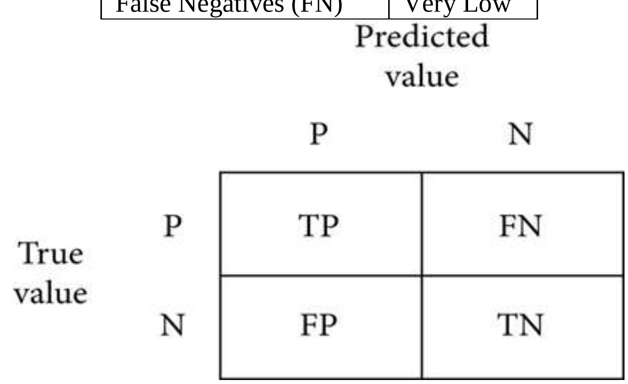


Fig 20 Confusion Matrix

A low false-negative rate is particularly important, as missing a heart disease diagnosis can lead to severe health consequences.

Discussion of Findings

The experimental findings highlight several important aspects:

- Machine learning models provide reliable baseline performance
- KNN performs well when combined with proper normalization
- Ensemble models improve robustness but may increase complexity
- Deep learning captures complex nonlinear relationships more effectively
- Data preprocessing is a critical factor influencing model success

Overall, the results validate the effectiveness of the proposed HDPS and confirm that integrating preprocessing techniques with advanced learning models leads to improved prediction accuracy and clinical reliability [17].

VI. CONCLUSION

This research successfully designed and evaluated an intelligent Heart Disease Prediction System (HDPS) using machine learning and deep learning techniques to support early and accurate diagnosis of cardiovascular diseases. By utilizing standard clinical data and adopting a structured, multi-phase implementation strategy, the study systematically examined the influence of data preprocessing, feature selection, normalization, and outlier detection on predictive performance. The use of the UCI Cleveland Heart Disease dataset ensured consistency with established research while enabling meaningful comparison of different predictive models.

The experimental results clearly demonstrate that data quality plays a crucial role in medical prediction systems. Models trained on raw data showed limited and inconsistent performance, whereas those trained on fully preprocessed datasets exhibited significant improvements in accuracy, stability, and generalization. Among the machine learning approaches, the K-Nearest Neighbors algorithm achieved the best results, particularly after normalization and outlier handling, highlighting the effectiveness of similarity-based learning in structured clinical datasets. However, the deep learning model consistently outperformed all traditional machine learning models by capturing complex and nonlinear relationships among medical attributes, leading to superior prediction accuracy and sensitivity.

From a clinical perspective, the proposed HDPS shows strong potential as a decision-support tool that can assist healthcare professionals in identifying high-risk patients at an early stage. High sensitivity and low false-negative rates observed in the best-performing models are especially important in medical diagnostics, where missed diagnoses can have severe consequences. By reducing reliance on manual evaluation and subjective judgment, the system enhances diagnostic consistency, efficiency, and reliability.

Overall, this research confirms that machine learning and deep learning techniques can significantly improve heart disease prediction and support preventive healthcare initiatives. While the current study is limited by the size and scope of the dataset, it provides a solid foundation for future research. Further work can focus on incorporating larger and more diverse datasets, integrating real-time patient monitoring data, and exploring hybrid or explainable AI models to enhance transparency and clinical adoption. The findings contribute meaningfully to the growing field of AI-driven healthcare and highlight the potential of intelligent systems in improving patient outcomes and healthcare delivery.

REFERENCES

1. World Health Organization, Cardiovascular Diseases (CVDs), WHO Fact Sheets, 2023.
2. D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019. [Online]. Available: https://archive.ics.uci.edu/ml
3. K. Polat, S. Şahan, and S. Güneş, “Automatic Detection of Heart Disease Using an Artificial Immune Recognition System,” Pattern

Recognition, vol. 39, no. 11, pp. 2186–2197, 2006.
4. N. R. Al-Milli, "Backpropagation Neural Network for Prediction of Heart Disease," Journal of Theoretical and Applied Information Technology, vol. 56, no. 1, pp. 131–135, 2013.
5. T. R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm," International Journal of Computer Science and Applications, vol. 6, no. 2, pp. 256–261, 2013.
6. S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249–268, 2007.
7. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
8. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
9. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
10. F. Chollet, Deep Learning with Python, Manning Publications, 2018.
11. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
12. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.
13. S. Dash, S. Shakyawar, M. Sharma, and S. Kaushik, "Big Data in Healthcare: Management, Analysis and Future Prospects," Journal of Big Data, vol. 6, no. 54, 2019.
14. A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," New England Journal of Medicine, vol. 380, pp. 1347–1358, 2019.
15. M. Nilashi, O. Ibrahim, and H. Ahmadi, "Accuracy Improvement for Heart Disease Prediction Using Machine Learning," Expert Systems with Applications, vol. 42, no. 5, pp. 2800–2812, 2015.
16. H. K. Shakya, "Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction," International Journal of Computer Applications, vol. 174, no. 8, pp. 12–18, 2017.
17. M. A. Khan, "An Intelligent Framework for Heart Disease Prediction Using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, pp. 204–210, 2020.