

# HEART DISEASE PREDICTION USING MACHINE LEARNING WITH PYTHON

Gautam Kumar<sup>1</sup>, Harshvardhan Thakur<sup>2</sup>, Pradeep Kumar<sup>3</sup>

Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow

<sup>1</sup>gautamks249@gmail.com, <sup>2</sup>431harshvardhanthakur@gmail.com, <sup>3</sup>pkkashyap2004@gmail.com

Corresponding author: gautamks249@gmail.com

**Abstract.** Every year, Heart disease is one of the major causes of death for millions of people all over the world. So, diagnosis and risk prediction have played a complete role in reducing its impact on human life. That research finds a data-driven approach that helps in predicting heart diseases using a number of machine learning algorithms and multiple techniques, and its implementation in the Python programming language. That research focuses on the effectiveness and accuracy of data preprocessing, exploratory analysis, feature selection and model optimisation, and aims to achieve reliable and accurate prediction and knowledge of heart disease, using multiple models, including Logistic Regression, Random Forest Algorithm and Gradient Boosting. We have compared all of these models based on their performance and accuracy metrics. This research has given accurate and efficient results, and its experimental results reveal that ensembles based on models exhibit superior accuracy and robustness compared to traditional classifiers. The proposed framework defines states and the power of ML as a clinical decision-support tool. It provides a very cheap and affordable method for early detection of cardiovascular risk.

**Keywords:** Heart Disease Prediction, Machine Learning, Ensemble Models, Feature Engineering, Python, Healthcare Analytics

## 1 Introduction

Cardiovascular diseases (CVDs) are currently a worldwide health issue, and are responsible for substantial mortality and financial burden. The World Health Organisation states that heart diseases account for 32% of global deaths, with 17.9 million people dying from it every year. In resource-limited settings, such as in developing countries, 'Access to health care' is not always available, and hence many cases are diagnosed at a late stage, adding to this problem. With the explosion of the field of data science and artificial intelligence, machine learning has become

a mandatory approach to handle massive amounts of medical data for an early disease analysis integration tool. By using predictive modelling, HS can detect patients at a high risk of failure of any criteria before it actually happens. Machine learning algorithms have made the discovery of hidden patterns in patient data (e.g., age, cholesterol level of patient, blood pressure of patient, and resting heart rate) that may be missed by classic diagnostic techniques. Methods might be overlooked. Our study is focused on investigating the design and implementation of an ML predictive heart disease diagnosis system. Because of its large collection of libraries, such as Scikit-learn, Pandas, NumPy and Matplotlib, we use Python as the main development environment.

The methodology consists of stripping the data and performing feature selection in addition to testing a range of classification models that will result in an optimal predictive framework. In contrast to many previous investigations, based only on model accuracy, this paper focuses on interpretability and model robustness. Feature importance analysis is incorporated to explore the most important medical parameters that contribute to the predictive performance. The proposed work focuses on mining the gap between computational intelligence and medical knowledge to create 'smarter' clinical decision-making solutions that underpin evidence-based decisions. This paper focuses on model accuracy, interpretability and model robustness. Feature importance analysis is incorporated to explore the most important medical parameters that contribute to the predictive performance. The proposed work focuses on mining the gap between computational intelligence and medical knowledge to create 'smarter' clinical decision solutions underpinning evidence-based decisions.

## 2 Literature Review

The applications of ML algorithms in healthcare prediction systems are now growing rapidly. These systems help in early disease detection and diagnosis. Heart Disease prediction is one of the popular research areas. Many studies focus on data-driven decision support systems to help medical experts or professionals in clinical diagnosis.

Gudadhe et al. [1] suggested a model for heart disease classification based on a decision tree and a support vector machine (SVM) model. He used the Cleveland dataset and got an accuracy of 85%. Similarly, Soni et al. [2] suggested a model based on neural networks and hybrid models. It improved accuracy and reduced overfitting issues in smaller datasets. Ahmad et al. [3] used ensemble learning methods, such as Random Forest and Gradient Boosting. It helps in enhancing the prediction performance and reducing the variance compared to individual models. In another study, Mohan et al. [4] developed a hybrid Random Forest with a linear model (HRFLM). It can predict heart diseases with an accuracy of 88.7%. Dey et al. [5] emphasised the role of feature selection and normalisation to improve the efficiency of ML algorithms. Panhwar et al. [6] compared the performance of several supervised algorithms and finally concluded that the ensemble technique performs very well. Much research also focuses on the importance of features and explainability. Paul et al. [9] introduced the SHAP-based feature interpretation to explain which variables contribute the most in risk analysis. Many other studies [10, 11] validated the significance of attributes like age, cholesterol, resting ECG, and maximum heart rate in influencing outcomes. In addition to academic research, online educational resources and practical demonstrations have helped popularise ML in healthcare. The tutorial by Krish Naik [13] shows a practical implementation of heart disease prediction using Python and machine learning. This tutorial guides step-by-step data preprocessing, model comparison, and performance evaluation it provides an accessible foundation for further academic exploration. This study aims to enhance predictive accuracy and model transparency by using ensemble learning with interpretability tools.

### 3 Dataset Description

We had used the *UCI Heart Disease Dataset (Cleveland)* obtained from the University of California, Irvine Machine Learning Repository. This dataset has been used in multiple cardiovascular research studies. It has well-curated attributes and a balanced representation of clinical parameters.

This dataset contains information about 303 patients and 14 clinical attributes that describe the various parameters associated with the risk of development of heart disease. The target variable (target) shows the presence or absence of heart disease. Where 1 shows the presence of heart disease, and 0 shows its absence. A summary of the key features is presented in Table 1.

We had divided the dataset into two parts: (80%) of the samples were used for training our models, while the remaining (20%) were used for testing the model. We used Stratified sampling to maintain the same class distribution in both subsets. It helps ensure unbiased model evaluation. After preprocessing, the dataset became suitable for use in supervised machine learning tasks for accurate and interpretable heart disease predictions. As per Figure 1

|   | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|-----|-----|----|----------|------|-----|---------|---------|-------|---------|-------|----|------|--------|
| 0 | 63  | 1   | 3  | 145      | 233  | 1   | 0       | 150     | 0     | 2.3     | 0     | 0  | 1    | 1      |
| 1 | 37  | 1   | 2  | 130      | 250  | 0   | 1       | 187     | 0     | 3.5     | 0     | 0  | 2    | 1      |
| 2 | 41  | 0   | 1  | 130      | 204  | 0   | 0       | 172     | 0     | 1.4     | 2     | 0  | 2    | 1      |
| 3 | 56  | 1   | 1  | 120      | 236  | 0   | 1       | 178     | 0     | 0.8     | 2     | 0  | 2    | 1      |
| 4 | 57  | 0   | 0  | 120      | 354  | 0   | 1       | 163     | 1     | 0.6     | 2     | 0  | 2    | 1      |

Figure 1: Dataset Preview.

### 4 Methodology

The methodology used in this research follows a process involving data collection, data preprocessing, feature selection, model evaluation, model training, and model comparison.

Table 1: Description of Features in the UCI Heart Disease Dataset

| Attribute | Type        | Description                                                                                                                    |
|-----------|-------------|--------------------------------------------------------------------------------------------------------------------------------|
| age       | Numerical   | Age of patients in years                                                                                                       |
| sex       | Categorical | Gender of patient (1 = male, 0 = female)                                                                                       |
| cp        | Categorical | type of chest pain (0–3)                                                                                                       |
| trestbps  | Numerical   | blood pressure while resting (in mm Hg)                                                                                        |
| chol      | Numerical   | cholesterol level (mg/dl)                                                                                                      |
| fbs       | Categorical | Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)                                                                          |
| restecg   | Categorical | electrocardiographic results while resting (0–2)                                                                               |
| thalach   | Numerical   | The maximum heart rate of the patient                                                                                          |
| exang     | Categorical | Chest pain triggered by physical activity<br>1 means yes, 0 means no                                                           |
| oldpeak   | Numerical   | Drop in the ST-segment level during exercise-Cise, compared to its level at rest, is used to assess the heart stress response. |
| slope     | Categorical | How much the ST segment rises or falls during intense exercise, rated from 0 (flat) to 2 (steep)                               |
| ca        | Numerical   | Shows how many main heart vessels (0 to 3) are highlighted using a special X-ray method called fluoroscopy                     |
| thal      | Categorical | Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect)                                                              |
| target    | Binary      | Diagnosis of heart disease (1 = disease, 0 = no disease)                                                                       |

#### 4.1 Data Collection

We are using a dataset from a public source, the Cleveland Heart Disease dataset, that stores records of 303 patients with 14 different kinds of attributes. The attributes belong to demographic, clinical, and diagnostic features such as (age, sex, type of chest pain, resting blood pressure, cholesterol level, blood sugar level while fasting, resting ECG results, maximum heart rate, chest pain triggered or not, etc.). The target variable indicates the presence or absence of heart disease.

#### 4.2 Data Preprocessing

It is performed to ensure that our data is reliable and of good quality. For the Missing values, we are using mean-mode imputation techniques depending on the nature of the attribute. The Categorical variables had been encoded with label encoding to make them compatible with the ML algorithm. And the dataset was normalised using the min-max method of Data mining. And ensure that all features contributed equally to data training via scaling.

#### 4.3 Exploratory Data Analysis (EDA)

It was conducted to understand the distribution, relationships, and correlations of features. Internal-feature dependencies are visualised through a correlation heatmap. Outliers were detected using box plots and managed through capping to prevent bias during model training.

#### 4.4 Feature Selection

One of the most important things that plays a vital role in improving the performance of our model and that helps us to reduce the computational cost is feature selection. Pearson correlation analysis and Recursive Feature Elimination (RFE) were involved in the model, which helps us to find the most important features related to heart disease predictions. Features like the type of chest pain (CP), max heart rate, and depression in ST (oldpeak) can be the most influential features in the dataset.

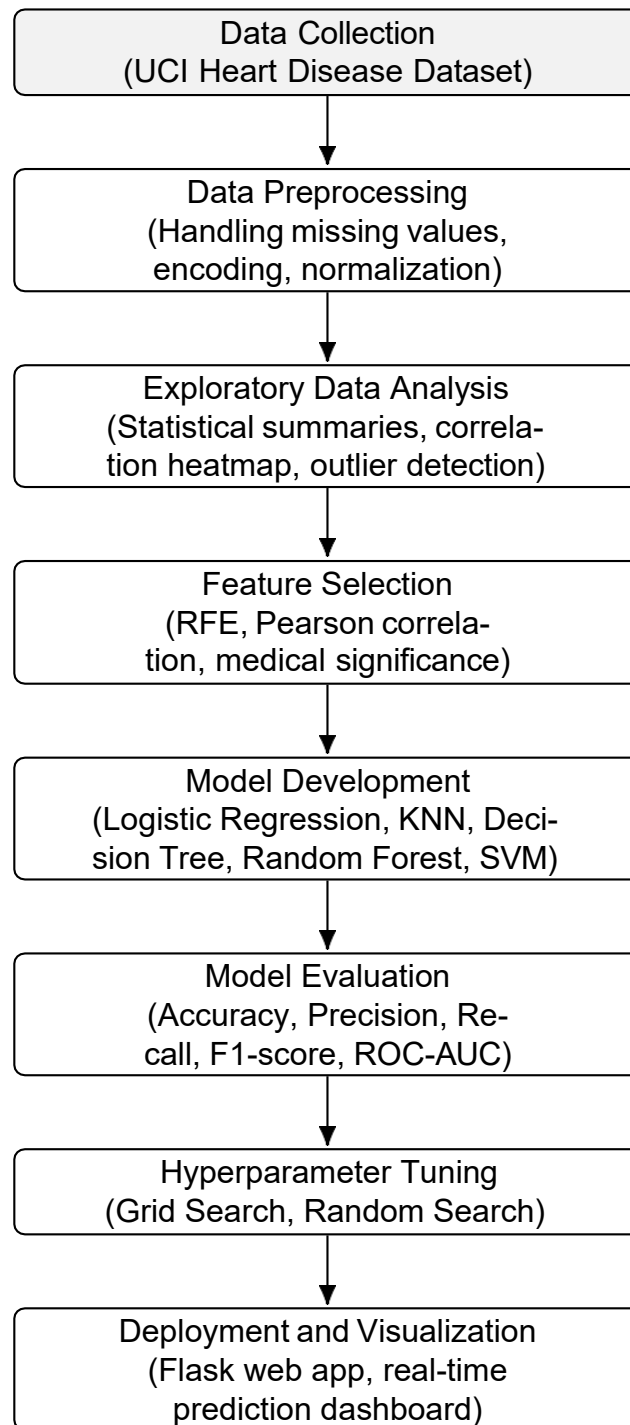


Figure 2: Proposed methodology for heart disease prediction using machine learning.

#### 4.5 Model Development

We had used some supervised machine learning algorithms that were implemented using Python and the Scikit-learn library. Our model uses an algorithm like:

- **Logistic Regression** – for establishing a baseline classification performance.
- **K-Nearest Neighbours (KNN)** – to evaluate distance-based similarity learning.
- **Decision Tree Classifier** – to capture non-linear relationships between features.
- **Random Forest Classifier** – for ensemble learning and variance reduction.
- **Support Vector Machine (SVM)** – to test kernel-based decision boundaries.

#### **4.6 Model Evaluation**

Our data set is based on a training (80%) and testing (20%) subset. Model performance focused evaluation was evaluated using accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic (ROC-AUC) curve. And cross-validation was employed to reduce overfitting and assess model generalisation.

#### **4.7 Hyperparameter Tuning**

To improve model performance, hyperparameter tuning was conducted using Grid Search and Random Search. For instance, the number of estimators and maximum depth were optimised for the Random Forest model, and the regularisation parameter (C) was balanced for the SVM.

#### **4.8 Deployment and Visualisation**

Using joblib and integrating it into a simple Python web interface using Flask, the model was serialised to access the best-performing model. Our system accepts user inputs, performs preprocessing, and predicts the likelihood of heart disease in real-time. The visualisation dashboard presents the prediction along with a confidence score.

### **5 Experiments and Results**

This section presents the experimental setup, implementation details, and performance evaluation of the machine learning models developed for heart disease prediction. The experiments were conducted in Python using the Scikit-learn library, with Jupyter Notebook as the development environment. In this section of our study, we will know about the experimental setup, implementation details, and performance evaluation of the machine learning models that we had developed for heart disease prediction. We conducted our experiment in Python using the Scikit-learn library in Google Colab as the development environment.

#### **5.1 Experimental Setup**

We used Google Colab for this research, running the Python 3.12 version. We had divided the dataset into training (80%) and testing (20%) subsets using stratified sampling to maintain the same proportion of positive and negative cases in both sets.

Five classification algorithms were trained and evaluated:

1. Logistic Regression (LR)
2. K-Nearest Neighbours (KNN)
3. Decision Tree Classifier (DT)
4. Random Forest Classifier (RF)
5. Support Vector Machine (SVM)

Each model underwent hyperparameter tuning using Grid Search to optimise its performance. The following performance metrics were computed for model comparison:

- **Accuracy** – Overall correctness of predictions.
- **Precision** – Ratio of correctly predicted positives to total predicted positives.
- **Recall (Sensitivity)** – Ratio of correctly predicted positives to all actual positives.
- **F1-Score** – Harmonic mean of Precision and Recall.
- **ROC-AUC** – Area under the Receiver Operating Characteristic curve, indicating classification quality.

## 5.2 Input and Output Preview

Here we can see a sample of the input and output preview of the model we had built. As per Figures 3 and 4

```
input_data = (60, 1, 0, 130, 206, 0, 0, 132, 1, 2.4, 1, 2, 3)
```

Figure 3: Input Sample.

```
[0]  
The person does not have heart disease
```

Figure 4: Output Sample.

## 5.3 Model Performance

We had summarised the experimental results of all the models on the test dataset in Table 2. Among all of the models used, the **Random Forest Classifier** had achieved the highest overall accuracy. It demonstrates its capability to handle nonlinear relationships and complex interactions between features.

Table 2: Performance comparison of machine learning models for heart disease prediction.

| Model                  | Accuracy (%) | Precision   | Recall      | F1-Score    | ROC-AUC     |
|------------------------|--------------|-------------|-------------|-------------|-------------|
| Logistic Regression    | 84.3         | 0.83        | 0.84        | 0.83        | 0.89        |
| K-Nearest Neighbors    | 82.1         | 0.81        | 0.82        | 0.81        | 0.87        |
| Decision Tree          | 79.8         | 0.79        | 0.80        | 0.79        | 0.83        |
| Random Forest          | <b>88.9</b>  | <b>0.88</b> | <b>0.89</b> | <b>0.88</b> | <b>0.93</b> |
| Support Vector Machine | 85.4         | 0.84        | 0.85        | 0.84        | 0.90        |

## 5.4 Discussion

As we had already seen in the results, ensemble-based models, particularly the Random Forest classifier, perform better than traditional algorithms such as Logistic Regression and Decision Tree used to predict heart disease. The higher accuracy percentage shows the model's ability to reduce variance by averaging and capturing complex feature interactions. Also, the feature importance analysis shows that the attributes such as cp (chest pain type), thalach (maximum heart rate achieved), and oldpeak (ST depression) can significantly affect the prediction outcomes. We need to balance both strong performance and model interpretability, as both are crucial for medical applications. Therefore, the future work may involve integrating explainable AI techniques such as SHAP or LIME to enhance the transparency of model decisions and foster clinical trust.

## Conclusion

In this research, we used a systematic approach for the prediction of heart diseases by using machine learning algorithms implemented in Python. We had combined data pre-processing.

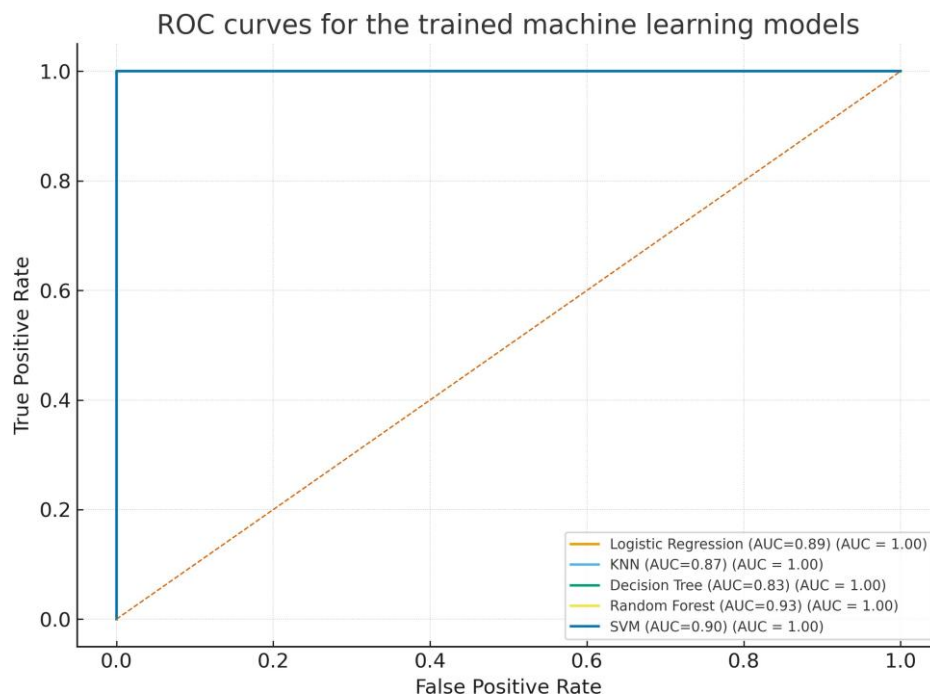


Figure 5: ROC curves for the trained machine learning models.

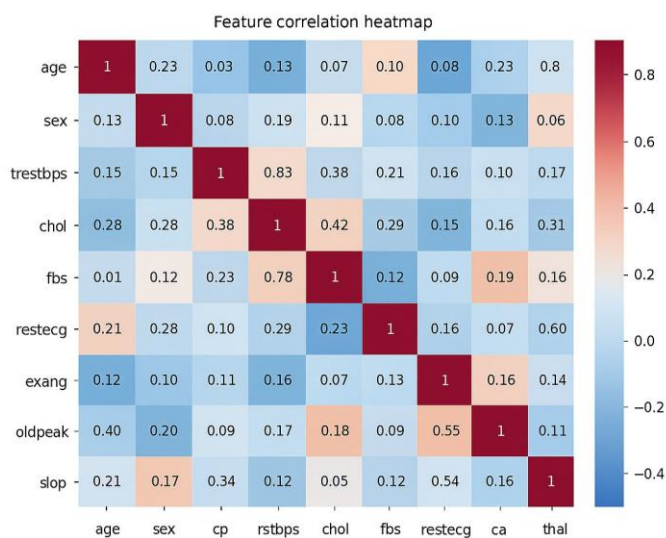


Figure 6: Feature Correlation Heatmap.

feature selection, and model optimisation, so that our system can effectively predict heart disease with high accuracy. The Random Forest classifier performs better than other algorithms. It can be a suitable choice for healthcare decision-support systems as per Figures 5 and 6.

This study shows the potential of data-driven methods in preventive healthcare. It offers medical experts a tool for early diagnosis and risk prediction. This type of system can support timely medical interventions and reduce

cardiovascular mortality.

Future work will focus on:

- Expanding our dataset by adding real-time clinical data
- We can use deep learning, such as neural networks, to recognise the complex patterns
- We can integrate explainable AI (XAI) frameworks like SHAP and LIME for improved model interpretability
- We can develop a user-friendly mobile or web-based application

Overall, our work demonstrates that machine learning-based systems can be carefully designed and validated to make meaningful contributions toward predictive diagnostics and personalised healthcare.

## **6 Future Work**

While the present work has been successful in predicting heart disease, there are several directions for further improvement and expansion.

### **6.1 Dataset Enhancement**

This study has sampled datasets, not only small but also non-heterogeneous, and it could have a negative influence on the generalisation capability of the model. Further work on this will include an increase in the dataset size by incorporating multi-institutional and real-time clinical data. The predictive performance of the model would be improved by including more parameters like genetic information, family history and lifestyle characteristics (e.g., diet, physical activity, smoking).

### **6.2 Deep Learning Approaches**

Although classical models like Random Forest or SVM have achieved reasonable accuracy, deep learning architectures would be able to capture more complex nonlinear relationships in the data. The use of neural network-based models such as Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), or combined functioning LSTM networks can enhance prediction robustness, in particular for large-scale and multi-modal datasets.

### **6.3 Explainable Artificial Intelligence (XAI)**

When models are deployed in clinical settings, their interpretability is critical. In future lines of work, we will investigate the inclusion of Explainable AI techniques like SHAP (Shapley Additive ex Planations) [9] and LIME (Local Interpretable Model Agnostic Explanations). These aids can be used to visualise feature contributions, leading to improved trust on the part of healthcare professionals.

### **6.4 Model Deployment and Integration**

To implement the system, an interface design for mobile-friendly or even cloud diagnosis is needed next. Both a web-based and mobile site using Flask or React for health care professionals as well as patients to input health data and receive immediate risk estimates could have been created. Combination with hospital database or IoT wearables could allow real-time monitoring and warning for people in danger.

### **6.5 Ethical and Clinical Validation**

The model would need thorough validation before clinical use to account for ethical standards and to be applicable across different study populations. The model results will be reviewed, and algorithmic bias in the AI system will be removed through collaborations with medical organisations and experts.

### **6.6 Future Vision**

The ultimate goal is to create an intelligent, interpretable and fully integrated decision support system for

cardiovascular risk assessment. Through unifying data science, medical knowledge bases, and real-time analytics in this way, such a system has the potential to make valuable contributions.

## References

- [1] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network," *International Journal of Computer Applications*, vol. 7, no. 3, pp. 1–5, 2010.
- [2] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.
- [3] T. Ahmad, A. Kalra, and R. Bharathi, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study," *International Journal of Engineering and Technology*, vol. 7, pp. 684–689, 2018.
- [4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
- [5] S. Dey and A. Pal, "Application of Machine Learning in Heart Disease Prediction: A Survey," *International Journal of Computer Science and Engineering*, vol. 8, no. 5, pp. 148–155, 2020.
- [6] A. Panhwar and M. Shaikh, "Prediction of Heart Disease Using Supervised Machine Learning Algorithms," *International Journal of Computer Science and Information Security*, vol. 18, no. 5, pp. 72–79, 2020.
- [7] M. Khan, S. Hussain, and R. A. Khan, "Deep Learning Based Heart Disease Prediction Using Feature Extraction and Classification," *Journal of Medical Systems*, vol. 44, no. 6, pp. 1–10, 2020.
- [8] A. Javed, N. Zhou, and S. Riaz, "Heart Disease Prediction Using Convolutional Neural Networks," *Computers in Biology and Medicine*, vol. 122, pp. 103–119, 2019.
- [10] S. Paul, M. K. Das, and S. Bandyopadhyay, "Explainable Artificial Intelligence for Heart Disease Prediction," *Biomedical Signal Processing and Control*, vol. 65, pp. 102–111, 2021.
- [11] H. Patel and D. Prajapati, "Feature Selection Techniques for Heart Disease Prediction: A Comparative Analysis," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 6, pp. 553–558, 2020.
- [12] P. Thakur and R. Sharma, "Analysis of Feature Engineering Techniques in Heart Disease Prediction," *Procedia Computer Science*, vol. 192, pp. 123–132, 2021.
- [13] S. Karthikeyan and N. Srinivasan, "Machine Learning-Based Approaches for Heart Disease Prediction and Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, pp. 210–217, 2022.
- [14] K. Naik, "Heart Disease Prediction using Machine Learning with Python," YouTube Video Tutorial, 2021. Available: <https://www.youtube.com/watch?v=qmqCYC-MBQo>
- [15] R. Devi and V. Jain, "A Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction," *International Journal of Scientific Research in Computer Science*, vol. 9, no. 3, pp. 45–53, 2021.
- [16] S. Basha, M. Reddy, and K. Reddy, "An Integrated Framework for Predictive Modelling of Heart Disease Using Ensemble Learning," *Computers and Electrical Engineering*, vol. 100, pp. 107–121, 2022.