WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Diseases Prediction

Sonam Soni [1], Abhay kumar Singh [2], Ujala Chinchkhede [3], Pragati Singh [4], Manoj Kumar Singh [5]

[1 2 3 4 5]Computer Science and Engineering Department, Shri Shankaracharya Technical Campus, Bhilai, Chhattisgarh, India

ABSTRACT :

This research presents the development of an automated heart disease prediction system employing three machine learning algorithms: K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression. To estimate the probability of heart disease based on key clinical features—age, sex, blood pressure, cholesterol concentration, and other relevant parameters—and to deliver a reliable decision-support tool for clinicians.

A curated dataset comprising the aforementioned variables was used to train and evaluate each model. Performance was assessed via accuracy, precision, recall, and F1-score metrics. KNN, Random Forest, and Logistic Regression were implemented and compared under identical data-preprocessing and validation protocols

KNN achieved the highest classification accuracy, while Random Forest and Logistic Regression also demonstrated strong predictive capabilities, albeit slightly lower. These findings validate the utility of machine learning—particularly KNN—for cardiovascular risk stratification. Logistic Regression's simplicity and interpretability clarify the association between patient features and disease likelihood. By identifying high-risk individuals, the system enables timely intervention, potentially reducing morbidity and mortality. Future work will focus on improving generalizability through larger, more diverse datasets, incorporating additional clinical parameters, and exploring ensemble strategies to enhance robustness and accuracy.

Keywords: Heart Disease, Machine Learning, KNN, Random Forest, Logistic Regression, Prediction, Healthcare

INTRODUCTION

Cardiovascular diseases, particularly heart disease, remain one of the leading causes of death globally. The early detection of heart-related conditions is essential to improve patient outcomes and reduce the burden on healthcare systems. However, conventional diagnostic techniques often depend on clinical assessments, which may not always provide timely or accurate predictions.

Machine learning (ML) has emerged as a promising tool in the medical field, offering data-driven approaches to assist in disease prediction and diagnosis ML models are capable of examining extensive patient data to identify patterns that may remain undetected through conventional methods. These models improve diagnostic accuracy, increase efficiency, and support evidence-based medical decisions.

The primary goal of this study is to build and evaluate a heart disease prediction system using three machine learning algorithms—K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression—based on essential clinical features. This model aims to support early diagnosis by identifying high-risk individuals based on medical indicators. It also seeks to assist healthcare professionals in delivering timely and informed treatment interventions.

LITERATURE REVIEW

The prediction of heart disease using machine learning techniques has attracted significant attention due to the potential for early diagnosis and intervention. Several studies have explored different algorithms and medical datasets, showcasing their effectiveness in predicting cardiovascular risks. This section reviews key approaches and methodologies used in heart disease prediction.

Machine Learning for Heart Disease Prediction

Machine learning algorithms have proven highly effective in predicting heart disease, leveraging various health indicators such as age, cholesterol, blood pressure, and other clinical metrics. Early research by Detrano et al. (1989) focused on statistical methods to assess heart disease risk. Over time, more complex algorithms have been employed, and datasets like the Cleveland Heart Disease dataset have become crucial in building prediction models. These datasets contain vital features such as age, gender, cholesterol, and ECG results, which serve as key predictors for heart disease.

Algorithms for Heart Disease Prediction

Several machine learning algorithms have been applied to heart disease prediction with promising results:

1. **K-Nearest Neighbours (KNN)**: KNN is a simple yet effective classification algorithm that predicts heart disease risk based on the proximity of data points in the feature space. Studies like Kuhn et al. (2013) have highlighted KNN's ability to classify patients by comparing them to similar cases in the dataset, yielding intuitive predictions with relatively high accuracy.
2. **Random Forest**: Random Forest, an ensemble learning algorithm, has gained popularity due to its ability to handle complex, large datasets without overfitting. It performs well in predicting heart disease risk by combining multiple decision trees to provide a robust prediction model
3. **Logistic Regression**: Logistic Regression is widely used for binary classification tasks, including heart disease prediction. It is widely used in medical data due to its simplicity and interpretability, providing clear insights into the relationship between health features and the likelihood of heart disease.

Dataset and Evaluation Techniques

Most studies in heart disease prediction utilize publicly available datasets like the Cleveland Heart Disease dataset, which includes critical patient information such as age, cholesterol levels, and blood pressure. The effectiveness of the predictive models is evaluated using metrics such as accuracy, precision, recall, and F1 score. Studies have consistently shown that using machine learning models on these datasets leads to high accuracy in classifying patients into risk categories, providing valuable insights for healthcare professionals .

METHODOLOGY

This section outlines the approach used to develop and evaluate a heart disease prediction system using machine learning algorithms. The methodology is divided into the following key components: (1) Data Collection and Preprocessing, (2) Model Development and Training, (3) Model Evaluation, and (4) Model Validation.

a. Data Collection and Preprocessing

To develop an effective heart disease prediction system, a comprehensive dataset containing key medical features of patients was used. The following steps were performed to ensure the quality and usability of the data:

Dataset Acquisition:

- The Cleveland Heart Disease dataset was utilized, which contains 303 patient records and 14 clinical attributes, such as age, sex, cholesterol levels, and blood pressure.

Data Cleaning:

- Missing values were addressed by imputation (using the median for continuous variables and the mode for categorical variables).
- Duplicate records were removed to maintain data integrity and avoid bias.

Feature Selection and Transformation:

- Relevant features like age, sex, cholesterol, and blood pressure were selected based on their significance to heart disease prediction.
- Categorical variables were one-hot encoded, and continuous variables were normalized to have zero mean and unit variance.

Data Partitioning:

1. The dataset was split into training and testing subsets in a 70-30 ratio, with stratified sampling to maintain class distribution across both sets.

b. Model Development and Training

Three machine learning algorithms were used to predict heart disease: K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression. The following steps were involved in model development:

1. **K-Nearest Neighbours (KNN):**
   - The optimal **k** value was determined by performing cross-validation with values from 1 to 25. The model with the highest F1-score was selected
2. **Random Forest:**
   - The Random Forest model was trained with 100 decision trees, with hyperparameters such as maximum depth and minimum samples per leaf optimized using grid search.
3. **Logistic Regression:**

- Logistic Regression was trained with L_2 regularization to prevent overfitting, and the regularization parameter (**C**) was optimized using cross-validation.

c. Model Evaluation

To assess the performance of each model, several evaluation metrics were used:

1. Performance Metrics:

- Accuracy, precision, recall, F1-score, and ROC-AUC were calculated for each model to assess prediction quality.

2. Cross-Validation:

- Five-fold cross-validation was applied on the training set to ensure generalization and avoid overfitting.

d. Model Validation and Testing

After training the models, the final evaluation was performed on the test set to assess their effectiveness in predicting heart disease.

1. The trained models were tested on the unseen test set, which was not used during the training phase, to evaluate their generalization ability.
2. Various performance metrics, including accuracy, precision, recall, and F1-score, were recorded to quantify the models' predictive capabilities.
3. The results of all three models were compared to determine which algorithm provided the best performance.
4. The KNN algorithm demonstrated the most optimal performance, exhibiting the best balance of accuracy, precision, and recall compared to Random Forest and Logistic Regression.
5. Based on these evaluations, the KNN model was selected as the most reliable model for heart disease prediction.

RESULT

The predictive performance of the heart disease prediction system was thoroughly evaluated using the held-out test set, which consisted of unseen data to ensure an unbiased assessment of the model's ability to generalize. The models were compared based on a variety of performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC, to determine their effectiveness in accurately classifying individuals at risk for heart disease.
These metrics were chosen to capture the model's ability not only to correctly identify positive cases (i.e., patients with heart disease) but also to minimize false positives and negatives, ensuring a balanced and reliable prediction. By evaluating these models on multiple criteria, we aimed to identify the algorithm that provides the best overall performance in stratifying cardiovascular risk, which is essential for early diagnosis and effective treatment planning in healthcare settings.

a. Quantitative Performance of Predictive Models

The table below summarizes each algorithm's performance:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
K-Nearest Neighbours	**0.92**	**0.90**	**0.94**	**0.92**	**0.96**
Random Forest	**0.89**	**0.88**	**0.90**	**0.89**	**0.93**
Logistic Regression	**0.87**	**0.85**	**0.88**	**0.86**	**0.91**

b. Comparative Analysis

1. **K-Nearest Neighbours** (KNN) exhibited the highest overall accuracy and F1-Score, indicating an excellent balance between precision and recall.
2. **Random Forest** achieved a strong ROC-AUC, demonstrating robust discrimination capability even with class imbalance.
3. **Logistic Regression** provided the greatest interpretability; its coefficients highlighted the most influential clinical features despite slightly lower accuracy.

c. Additional Validation

1. **ROC Curves:** Figure 1 presents the receiver operating characteristic curves for all three models, confirming KNN's superior true-positive versus false-positive trade-off.
2. **Confusion Matrices:** Figure 2 displays the confusion matrix for the KNN classifier, illustrating its low false-negative rate, which is critical for medical screening applications.

DISCUSSION

The results of this study show that integrating machine learning algorithms with healthcare data has the potential to provide an efficient and reliable heart disease prediction system. The performance of the K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression models was thoroughly evaluated, with KNN demonstrating the highest accuracy in predicting heart disease risk. This indicates that distance-based algorithms, like KNN, may be particularly effective in identifying subtle patterns in patient data that correlate with heart disease.

One of the main advantages of this approach is its interpretability, especially with Logistic Regression, which allows healthcare professionals to understand the impact of individual features, such as age, blood pressure, and cholesterol levels, on heart disease risk. Additionally, these models can be trained using readily available medical data, making the approach highly scalable and adaptable to various healthcare environments.

However, the models' performance was limited by the dataset used, which may not fully capture the complexity of heart disease risk factors. For future work, expanding the dataset to include more diverse patient demographics and medical histories could enhance model accuracy and generalization.

CONCLUSION

This study demonstrates the successful development of an automated heart disease prediction system using K-Nearest Neighbours, Random Forest, and Logistic Regression. By leveraging key clinical features—including age, sex, blood pressure, and cholesterol levels—the models achieved high predictive performance, with KNN yielding the best overall accuracy and F1-score

The combination of data preprocessing, systematic hyperparameter optimization, and rigorous validation ensured that the system generalizes well to unseen patient records. Importantly, the interpretability of Logistic Regression provides clinicians with clear insights into how specific health indicators contribute to disease risk, enhancing trust and transparency. This work underscores the potential of machine learning to support early diagnosis and improve patient outcomes.

Future efforts will focus on incorporating larger, more heterogeneous datasets, exploring ensemble and hybrid modelling approaches, and integrating the system into clinical decision-support platforms to facilitate real-world deployment and continual model refinement.

REFERENCES :

[1] **Heart Disease Prediction Using ML** – Logistic Regression, Random Forest, KNN; feature selection and evaluation. *(IEEE Xplore)*

[2] **Comparative ML for Heart Disease** – Naive Bayes, Decision Trees, KNN; accuracy and recall. *(Springer)*

[3] **ML in Heart Disease Diagnosis** – Review of ML applications in clinical decision-making. *(PubMed Central)*

[4] **Heart Disease Datasets** – Overview and future directions. *(Frontiers)*

[5] **Feature Importance Analysis** – Key clinical predictors via permutation importance. *(ResearchGate)*