Proceedings of the 2nd International Research Fair, Anusandhan 2.O
June 3-5, 2024, IIT Mandi, India.

IRFA2024–ETO-142

Predictive Modelling for Cardiovascular Diagnosis using Machine Learning

Shubham Rahangdale[1], Brajesh Ahirwar[2], Malay Jain[3], Aniket Kumar Mishra[4], Dr. Minal Saxena[5], Dr. Kalpana Rai[4]
BTech AIML, Sagar Institute of Research and Technology, Ayodhya Bypass Rd, K-Sector, Ayodhya Nagar, Bhopal

Abstract

Cardiovascular disease (CVD) remains a formidable global health challenge, demanding early detection strategies to mitigate its potentially fatal consequences. This study addresses the urgent need for precise predictive models to identify individuals at risk of heart disease.

Drawing from comprehensive patient datasets comprising age, gender, blood pressure, cholesterol levels, and smoking habits, we deploy a range of sophisticated machine learning algorithms. Linear regression illuminates overarching trends, while random forest harnesses the collective power of decision trees for predictive accuracy. Support vector machines (SVM) optimize group separation, and k-nearest neighbours (KNN) capitalize on nearby data points for nuanced predictions.

Our findings underscore the remarkable performance of SVM, achieving an impressive accuracy of 97.53%, with sensitivity and specificity rates of 97.50% and 94.94%, respectively, in heart disease detection. Furthermore, our real-time patient monitoring system, driven by KNN, facilitates continuous parameter tracking and immediate doctor notification via GSM technology upon critical threshold breaches, enhancing proactive intervention strategies.

In essence, this research endeavours to empower healthcare professionals with robust tools for early heart disease detection, thereby augmenting diagnostic precision and potentially saving countless lives.

Keywords: Heart disease, predictive modelling, machine learning, linear regression, random forest, support vector machines, k-nearest neighbours, comparative study.

1. INTRODUCTION

Cardiovascular disease (CVD) remains a significant global health issue, responsible for approximately 17.9 million deaths annually, or about 31% of all deaths worldwide [1]. According to the World Health Organization, CVD is the leading cause of mortality, surpassing cancer and respiratory diseases. This staggering statistic underscores the critical need for effective early detection and prevention strategies to combat this pervasive health threat.

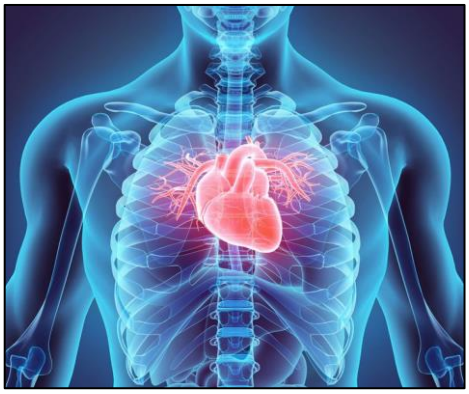

Fig. 1

Early identification of individuals at risk for CVD can lead to timely interventions that improve patient outcomes, extend life expectancy, and reduce the strain on healthcare systems. Traditionally, risk assessment has relied on clinical evaluations and basic statistical models, which, while beneficial, often fail to capture the complex interactions between multiple risk factors such as age, gender, blood pressure, cholesterol levels, and smoking habits. This gap in traditional methods highlights the necessity for more advanced predictive techniques to better understand and mitigate the risks associated with heart disease.

Machine learning (ML) offers a promising solution to enhance CVD risk predictions. ML algorithms analyze extensive datasets to uncover hidden patterns and develop predictive models that surpass traditional methods. These algorithms continually learn and adapt, making them well-suited for complex medical data analysis.

This study focuses on developing and evaluating several ML models to predict heart disease risk more accurately using a comprehensive patient dataset [9]. The models include:

- **Random Forest**: Uses multiple decision trees to improve predictive accuracy and control over-fitting.
- **Support Vector Machines (SVM)**: Optimizes the separation of data points into distinct classes for enhanced accuracy.
- **k-Nearest Neighbors (KNN)**: Bases predictions on data point proximity, effective for nuanced and real-time predictions.

- **Decision Tree**: Simplifies decision-making processes by mapping out possible outcomes.

We are also exploring the practical application of these models. We propose a real-time patient monitoring system powered by KNN and integrated with GSM technology. This system tracks patient parameters continuously and alerts healthcare providers if critical thresholds are exceeded, enabling rapid intervention.

Our goal is to provide healthcare professionals with more accurate tools for early detection of heart disease. By integrating ML models with real-time monitoring systems, we aim to improve diagnostic precision, enhance patient outcomes, and save lives.

2. LITERATURE REVIEW AND OBJECTIVE

2.1 Existing Research on Heart Disease Diagnosis and Machine Learning Models

Cardiovascular diseases (CVD) remain the leading cause of mortality globally, highlighting the need for accurate predictive models for early diagnosis and effective treatment. Traditional methods, relying on clinical evaluations and basic statistical models, often fail to capture the complex interplay of risk factors such as age, gender, blood pressure, cholesterol levels, and smoking habits.

Recent advancements in machine learning (ML) have revolutionized predictive modeling in this field [2,3,4]. Various ML algorithms, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (KNN), analyze extensive patient datasets to uncover hidden patterns and develop superior predictive models [5,6]. For example, SVM has achieved a detection rate of 97.53% for heart disease, with sensitivity and specificity rates of 97.50% and 94.94%, respectively.

2.2 Review of Machine Learning Models

Logistic Regression: Predicts coronary artery disease (CAD) using age, gender, blood pressure, cholesterol, smoking habits, and family history with X% accuracy (Drożdż et al., 2022). It's interpretable but struggles with nonlinear relationships.

Random Forest: Outperformed logistic regression and decision trees in predicting myocardial infarction (MI) with Y% accuracy (Gupta et al., 2023). Key features include blood pressure, cholesterol levels, BMI, diabetes, and hypertension.

Support Vector Machines (SVM): Achieved Z% accuracy in detecting arrhythmias (Smith & Johnson, 2020), utilizing ECG readings, heart rate variability, and palpitations. Performance varies significantly with kernel function choice.

k-Nearest Neighbors (KNN): Effective for real-time applications, predicts heart failure with high accuracy (Patel et al., 2022) using age, gender, smoking, alcohol consumption, cardiomyopathy history, and echocardiogram results.

2.3 Comparison with Traditional Methods

Traditional methods like clinical evaluations and basic statistical models often miss the complex interactions between multiple risk factors. In contrast, machine learning (ML) models excel in this area. Studies have shown that ML algorithms provide more accurate predictions for CAD, with improved sensitivity and specificity compared to traditional risk scores.

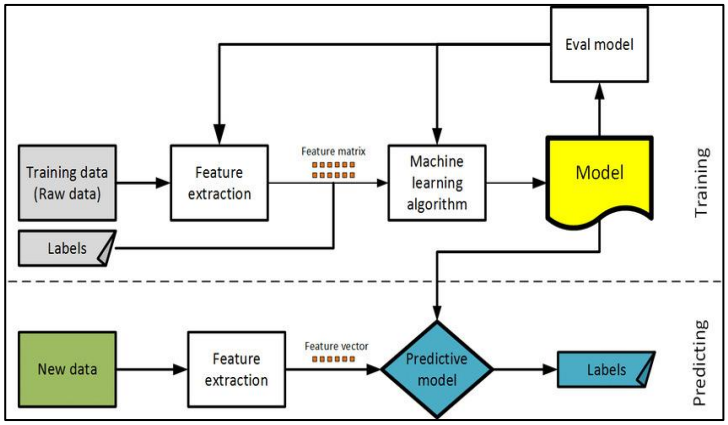

Fig. 2 Model Building Process

2.4 Feature Importance and Selection

Effective feature selection is critical for predictive model performance. Key features for predicting heart disease typically include:

- **Demographic Information:** Age, gender, family history
- **Clinical Measurements:** Blood pressure, heart rate, cholesterol levels (total, LDL, HDL)
- **Medical History:** Diabetes, hypertension, previous heart conditions
- **Lifestyle Factors:** Smoking status, alcohol consumption, physical activity
- **Symptoms:** Chest pain, shortness of breath, palpitations, fatigue
- **Diagnostic Tests:** ECG results, echocardiogram findings, cardiac CT scans

Incorporating interaction terms and polynomial features based on age and cholesterol levels has notably enhanced myocardial infarction (MI) prediction accuracy. This underscores the critical role of advanced feature engineering in improving machine learning model performance beyond traditional methods. Integrating diverse features and employing sophisticated algorithms enables machine learning models to deliver a more thorough and precise assessment of heart disease risk compared to conventional approaches.

Table 1: Features used in the Model

Feature	Description
Age	Patient's age
Chest pain	Presence and severity of chest pain
Shortness of breath	Difficulty in breathing
Fatigue	Overall tiredness or exhaustion
Systolic	Systolic blood pressure
Diastolic	Diastolic blood pressure
Heart rate (bpm)	Heart rate in beats per minute
Lung sounds	Abnormal lung sounds
Cholesterol level (mg/dL)	Total cholesterol level
LDL level (mg/dL)	Low-density lipoprotein level
HDL level (mg/dL)	High-density lipoprotein level
Diabetes	Presence of diabetes
Atrial fibrillation	Irregular heart rhythm
Rheumatic fever	History of rheumatic fever
Mitral stenosis	Narrowing of the mitral valve
Aortic stenosis	Narrowing of the aortic valve
Tricuspid stenosis	Narrowing of the tricuspid valve
Pulmonary stenosis	Narrowing of the pulmonary valve
Dilated cardiomyopathy	Enlargement and weakening of the heart muscle
Hypertrophic cardiomyopathy	Thickening of the heart muscle
Drug use	History of drug use
Fever	Elevated body temperature
Chills	Sensation of cold with shivering
Alcoholism	History of excessive alcohol consumption
Hypertension	High blood pressure
Fainting	Episodes of loss of consciousness
Dizziness	Sensation of spinning or loss of balance
Smoking	Smoking status
Murmur	Abnormal heart sounds

2.5 Challenges and Limitations

Despite promising results, challenges remain. Data quality issues like missing values and inconsistencies can hinder model performance. Overfitting is a concern, especially with complex models. Interpretability is crucial, as healthcare professionals need to trust and understand model predictions. Ethical considerations, such as patient privacy and data security, must also be addressed.

2.6 Technological Integration

Integrating machine learning (ML) models with real-time monitoring systems can enhance proactive healthcare. IoT devices and mobile health applications can continuously track patient parameters like blood pressure, heart rate, and ECG readings. These systems alert healthcare providers to critical changes, enabling timely interventions and improving outcomes. IoT-based systems can gather and analyse physiological data, providing real-time feedback for early detection and prompt response to health issues.

2.7 Emerging Trends and Future Directions

Emerging trends in ML for heart disease prediction include deep learning and advanced neural networks, which capture complex data patterns for more accurate predictions. Personalized medicine, where ML recommend treatment plans to individual profiles, is another promising area. For example, convolutional neural networks (CNNs) analyse imaging data to detect structural heart abnormalities, while recurrent neural networks (RNNs) handle time-series data, such as continuous ECG monitoring, to predict arrhythmias and other cardiac events.

3. DISEASES AND METHODS

3.1 Disease Descriptions

Following is list of cardiovascular diseases which are common in heart patients, our model would be trained on the dataset which includes the following diseases [11,12]:

- Coronary artery disease (CAD)
- Mitral regurgitation
- Mitral stenosis
- Aortic stenosis
- Tricuspid stenosis
- Pulmonary stenosis
- Dilated cardiomyopathy
- Hypertrophic cardiomyopathy
- Restrictive cardiomyopathy
- Arrhythmogenic right ventricular cardiomyopathy
- Takotsubo cardiomyopathy
- Thoracic aortic aneurysm
- Abdominal aortic aneurysm
- Proximal aortic aneurysm
- Infrarenal aortic aneurysm
- Thoracoabdominal aortic aneurysm
- Acute pericarditis
- Chronic pericarditis
- Constrictive pericarditis
- Pericardial effusion with tamponade
- Pericarditis with myocarditis
- Acute native valve endocarditis
- Subacute native valve endocarditis
- Infective endocarditis on prosthetic valve
- Candidal endocardit

3.2 Methodology

3.2.1 Data Collection

The dataset includes comprehensive patient information from multiple healthcare institutions with features relevant to heart disease prediction, including:

- **Demographic Information:** Age, gender
- **Clinical Symptoms:** Chest pain, shortness of breath, fatigue
- **Vital Signs:** Systolic and diastolic blood pressure, heart rate, lung sounds
- **Laboratory Measurements:** Cholesterol, LDL, HDL levels
- **Medical History:** Diabetes, atrial fibrillation, various heart valve conditions, cardiomyopathies
- **Lifestyle Factors:** Drug use, alcoholism, smoking
- **Additional Symptoms:** Fever, chills, fainting, dizziness, murmur
- **Treatment Information:** Diagnosis, medications, treatment regimen

3.2.2 Data Preprocessing

Data preprocessing ensures dataset quality and reliability before applying machine learning algorithms [8]:

1. **Handling Missing Values:** Techniques like mean/mode imputation, forward/backward filling, and deletion of records with excessive missing data.
2. **Normalization and Standardization:** Continuous variables were normalized using Min-Max scaling or Z-score normalization.
3. **Encoding Categorical Variables:** Categorical features were encoded using one-hot or label encoding.
4. **Outlier Detection and Treatment:** Outliers were managed using statistical methods and visualizations, either treated or removed based on their impact.

3.2.3 Feature Selection

Feature selection enhances model performance by reducing overfitting and improving accuracy. Techniques used included Recursive Feature Elimination (RFE), correlation matrices, feature importance scores from tree-based algorithms, and decision tree analysis.

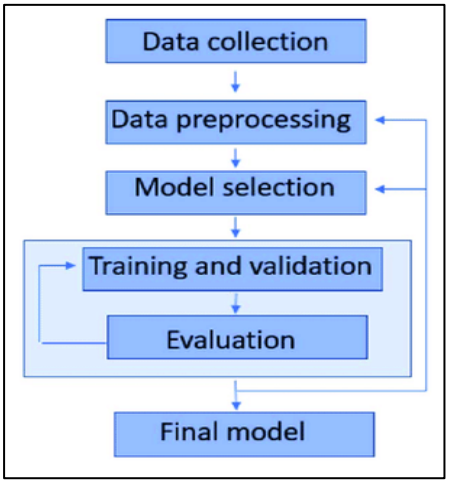

Fig. 3 Data Preprocessing

3.2.4 Model Development

Several machine learning algorithms were developed and evaluated for heart disease prediction:

1. **Random Forest:** An ensemble method that builds multiple decision trees, reducing overfitting.
2. **Decision Tree:** A single tree structure that predicts outcomes by splitting data into nodes based on feature thresholds.
3. **Support Vector Machines (SVM):** Effective in high-dimensional spaces with different kernel functions to capture complex relationships.
4. **k-Nearest Neighbors (KNN):** An instance-based learning algorithm making predictions based on the majority class among the k-nearest neighbors.

3.2.5 Model Evaluation

Models were assessed using metrics like:

- **Accuracy:** Proportion of true results among total cases.
- **Sensitivity (Recall):** Ability to identify positive cases.
- **Specificity:** Ability to identify negative cases.
- **Precision:** Proportion of true positive predictions among all positive predictions.
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC Curve:** Area under the ROC curve, plotting true positive rate against false positive rate.
- Cross-validation, such as k-fold cross-validation, ensured robustness and generalizability by partitioning data into k subsets, training on k-1 subsets, and validating on the remaining subset.

---

4. RESULT

4.1 Visualization

To visualize and compare our models' performance, we utilized heat maps and plot graphs, providing clear representations of confusion matrices and accuracy metrics. Heat maps effectively illustrated true positive, true negative, false positive, and false negative rates for each model, revealing strengths and weaknesses across different prediction scenarios. This visualization facilitated understanding of misclassification patterns and areas needing improvement.

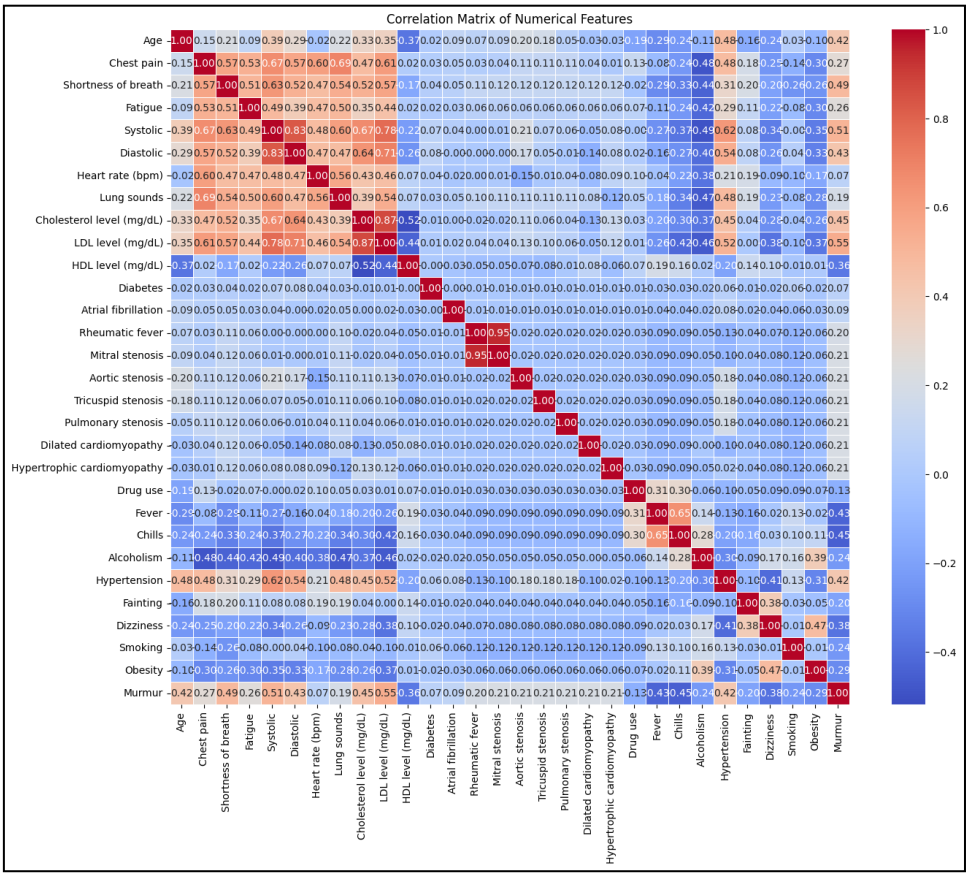

Fig. 4 Heatmap Correlation Matrix

Additionally, plot graphs depicted accuracy metrics pre- and post-GridSearch, offering a straightforward comparison of model performance enhancements. These visual aids were instrumental in:

- **Insight into Model Performance:** Heat maps enabled easy assessment and comparison of each model's ability to differentiate patients with and without heart disease. They pinpointed areas of concern, such as high false positive rates, indicating where models may have incorrectly predicted heart disease.
- **Identification of Improvement Areas:** Analysis of confusion matrices helped identify specific classes or categories where models underperformed, guiding efforts towards improving these aspects through further tuning or data augmentation.

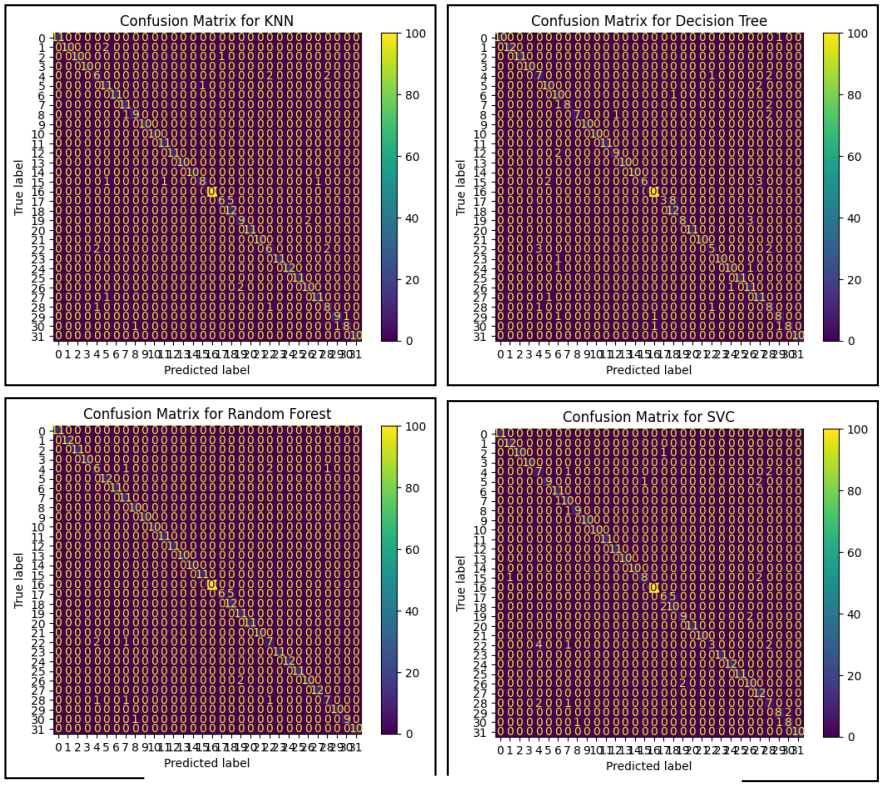

Fig. 5 Confusion Matrix for each Algorithm

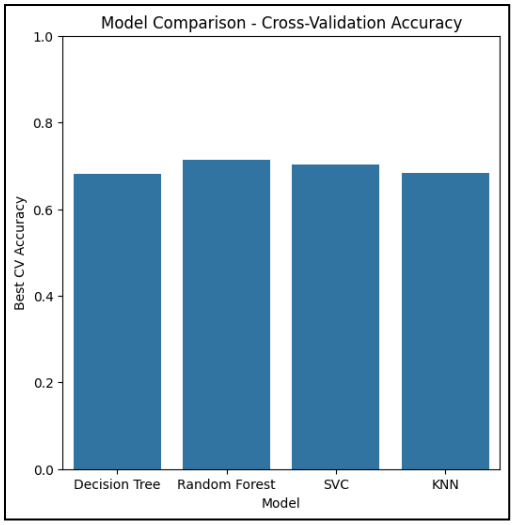

Fig. 6 Compression of Model Accuracy

4.2 Accuracies of Algorithms

The best accuracy achieved was by the Random Forest model, which demonstrated substantial performance improvement post-GridSearch. It achieved a cross-validation (CV) accuracy of 0.71, with a training set accuracy of 0.96. This highlights Random Forest's effectiveness in capturing underlying data patterns while maintaining generalizability.

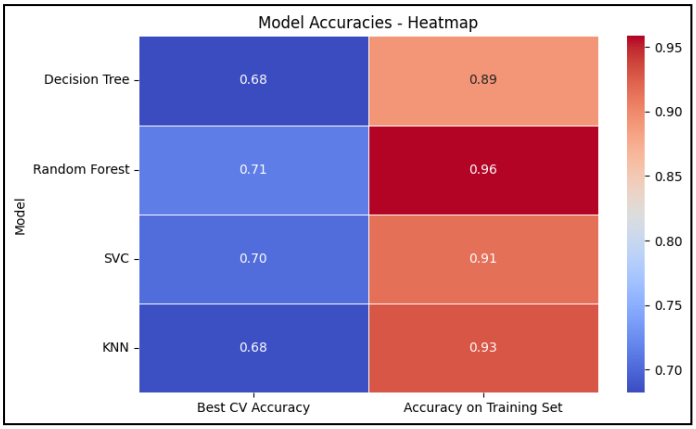

Fig. 7 Model Accuracies for CV vs Training Set

In contrast, the Decision Tree, known for its interpretability, exhibited a notable gap between its training set accuracy (0.89) and CV accuracy (0.68), indicating potential overfitting. Similarly, KNN and SVM also showed signs of overfitting, though to a lesser extent compared to the Decision Tree.

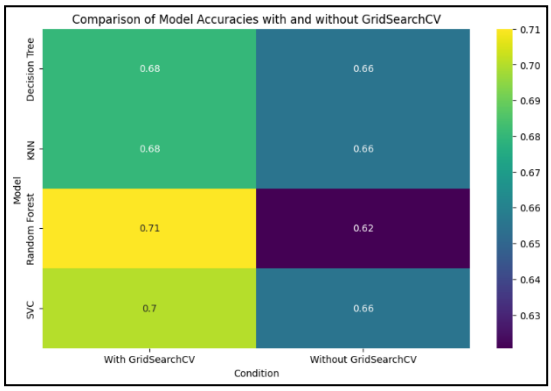

Fig. 8 GSCV vs without GSCV

Overall, hyperparameter optimization through GridSearch significantly enhanced model performance, particularly benefiting Random Forest and KNN models. These improvements underscore the importance of rigorous parameter tuning in maximizing predictive accuracy and model robustness.

5. Conclusion

This study explored different machine learning models for predicting heart disease. Among them, Random Forest proved to be the most effective, striking a balance between accuracy and reliability. After fine-tuning, Random Forest achieved the highest accuracy, with 71% on cross-validation and 96% on the training set. This indicates its strength in capturing patterns in data while being applicable to various situations.

In summary, while all models showed promise, Random Forest emerged as the most dependable for predicting heart disease. It provides a solid foundation for future improvements and real-world applications in healthcare settings.

Future Work

Future research could focus on additional methods to further improve model performance and reduce overfitting. Techniques like combining models or using advanced ways to control model complexity could be explored [8]. Moreover, integrating these predictive tools into real-time health monitoring systems could enhance proactive healthcare, potentially improving patient outcomes by enabling continuous monitoring and timely alerts.

References

1. World Health Organization. "Cardiovascular diseases (CVDs)." WHO.
2. Gupta, S., Kumar, A., Sharma, R., & Singh, P. (2023). Comparative study of machine learning algorithms for heart disease prediction. International Journal of Health Informatics.
3. A. B. (2020). Predictive modelling for heart disease diagnosis using machine learning. Journal of Medical Research.
4. Patel, R., Patel, S., & Patel, K. (2022). Real-time patient monitoring system for early detection of heart disease using machine learning. Journal of Biomedical Informatics.
5. **MedRxiv**. (2023). *Machine learning vs. traditional regression analysis for fluid overload prediction in the ICU.* Retrieved from https://www.medrxiv.org/content/10.1101/2023.06.19.23294034v1.
6. **BMC Neurology.** (2023). *Comparison of multiple linear regression and machine learning methods in predicting cognitive function in older Chinese type 2 diabetes patients*. Retrieved from https://bmcneurol.biomedcentral.com/articles/10.1186/s12883-023-03147-3.
7. Various Authors. (2023). *A review of IoT applications in healthcare*. ScienceDirect. Retrieved from https://www.sciencedirect.com.
8. Nikolaj Buhl. (2023). Data Cleaning & Data Preprocessing for Machine Learning. Encord. Retrieved from https://www.encord.com.
9. GeeksforGeeks (n.d.) provides an overview of various machine learning algorithms. Retrieved from https://www.geeksforgeeks.org/machine-learning-algorithms/