**Using Data mining techniques for discovering crucial features in predicting heart diseases**

Jaisaiarun P Srinivasan – 122003096

Anandalakshmi R – 122003023

Bhargavi T – 122003050

Cardiovascular diseases (CVDs) are one of the important causes of death worldwide. According to the Centers for Disease Control and Prevention (CDC), a heart attack occurs every 40 seconds in the United States. Prediction of CVDs plays a crucial role in the healthcare sector. Data mining techniques are used be operate upon the massive set of data collected from the healthcare sector to produce impeccable results in the prediction of CVDs. As mentioned, this combination has been a crucial topic of research in clinical data analysis. Unlike previous studies, in this paper, they have concentrated on finding the vital features using data mining for the prediction of CVDs.

In this paper, the combination of features is selected from a set of features and used with the classification techniques. The selection process is of the brute force approach. Three or more features are present in any combination set. After obtaining the evaluation metric for all the combinations for the classification models, the best feature set and the corresponding model is validated using a different dataset. Classification models considered are k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Vote (a hybrid technique with Naïve Bayes and Logistic Regression). Also, the dataset is taken from the UCI machine learning repository. The Cleveland database is used to identify the crucial features and the top-performing models. UCI Statlog Heart Disease dataset is utilized for validation of an identified set of crucial features and the model. The performance metrics used are accuracy, f-measure, and precision.

This study identifies nine significant features and three techniques that produce accurate results. The reliability of the results is confirmed through evaluation. The models obtained from this paper for predicting CVDs have an accuracy of 87.4%. The models used for prediction are enhanced with the help of these efficient and significant features obtained through data mining techniques.

**Keywords:** Cardiovascular diseases (CVDs), Data mining, Prediction model, Feature selection

# Base Paper Title: Identification of significant features and data mining techniques in predicting heart disease

**Base Paper URL:** <https://www.sciencedirect.com/science/article/abs/pii/S0736585318308876>



1. **Joy Christy**