International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Smart Health Monitoring System: Heart Attack Risk Prediction using Machine Learning

Suriya V, Tamilarasan R
Department of Computer Science and Information Technology
Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India
Under the guidance of
Dr. K. Sheela
Assistant Professor, Dept. of CS & IT, VISTAS

https://doi.org/10.55041/ijsmt.v2i5.017

Cite this Article: V, S. & R, T. (2026). Smart Health Monitoring System: Heart Attack Risk Prediction using Machine Learning. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.017

ABSTRACT

Abstract—Heart disease remains one of the leading causes of mortality worldwide. Early prediction of heart attack risk is crucial for preventive healthcare. This paper presents a machine learning-based system to predict the risk of heart attack using patient health and lifestyle data. A dataset of 8,763 records with 26 attributes was employed, encompassing factors such as age, cholesterol, blood pressure, heart rate, diabetes, smoking, obesity, exercise hours, stress levels, and BMI. Extensive data preprocessing, feature selection using the chi-square test, and three classification algorithms—Logistic Regression, Decision Tree, and Random Forest—were implemented and compared. Logistic Regression achieved the highest accuracy of 65.14%, followed by Random Forest at 60.69% and Decision Tree at 53.90%. A prediction function was developed to classify new patient data as either high or low risk in real time. The results demonstrate the potential of machine learning in supporting early cardiac risk assessment.

Keywords—Heart attack prediction; machine learning; logistic regression; random forest; decision tree; feature selection; healthcare analytics

I. INTRODUCTION

Cardiovascular diseases, particularly heart attacks, represent one of the foremost public health challenges globally. The World Health Organization estimates that millions of lives are lost annually due to cardiac events that could have been prevented with timely intervention. Lifestyle-related factors such as sedentary behavior, poor diet, smoking, stress, and obesity have accelerated the prevalence of these conditions.

Traditional clinical diagnosis of heart attack risk relies on physician expertise, medical history analysis, and laboratory test results. While effective, this approach is inherently time-consuming, resource-intensive, and subject to variability in accuracy depending on the clinician's experience and the completeness of available patient data. In environments where healthcare professionals are scarce—particularly in rural or under-resourced regions—timely diagnosis is even more challenging.

Machine learning offers a transformative approach to address these limitations. By training on large volumes of historical patient data, machine learning algorithms can identify complex patterns among health parameters that may not be apparent to human observers. This enables automated, data-driven predictions that can augment clinical decision-making.

This paper describes the development of an automated heart attack risk prediction system. The system accepts patient health data as input, applies preprocessing and feature selection, and outputs a risk classification using three supervised machine learning models. The objectives are: (1) to identify the most relevant risk-contributing features from health and lifestyle data; (2) to train and compare multiple classifiers; and (3) to provide a practical prediction interface for real-time use in clinical or personal health monitoring contexts.

II. RELATED WORK

A substantial body of literature exists on the application of machine learning to cardiovascular disease prediction. Géron [1] provides a foundational overview of machine learning techniques including ensemble methods applicable to medical prediction tasks. Mitchell [2] lays the theoretical groundwork for supervised learning algorithms used in classification problems.

Han et al. [4] demonstrate data mining concepts applicable to healthcare datasets, including handling of class imbalance and high-dimensional feature spaces. Several studies on the UCI Heart Disease Dataset [11] have applied logistic regression and tree-based methods to cardiac classification, reporting accuracy ranges of 60–85% depending on dataset size and feature engineering. Kaggle benchmark studies [10] further confirm that ensemble models such as Random Forest often outperform single classifiers when data is sufficient and well-preprocessed.

The present work differentiates itself by applying a comprehensive pipeline—including chi-square feature selection, mean/mode imputation, and standard scaling—on a diverse, multi-national dataset of 8,763 records, while maintaining focus on system practicality through an integrated prediction function.

III. DATASET AND FEATURE DESCRIPTION

The dataset used in this study consists of 8,763 patient records with 26 attributes. Each record captures a broad range of clinical and lifestyle parameters relevant to cardiovascular health. The target variable, "Heart Attack Risk," is a binary classification label indicating high or low risk.

A. Numerical Features

- Age
- Cholesterol level
- Resting heart rate
- Exercise hours per week
- Blood pressure (systolic)
- BMI (Body Mass Index)
- Stress level

B. Categorical Features

- Gender
- Diabetes (Yes/No)
- Smoking status
- Obesity
- Alcohol consumption
- Country of origin
- Family history of heart disease

The dataset exhibits class imbalance, with a higher proportion of low-risk cases. This affects model performance, particularly recall for the high-risk class, and is identified as an area for future improvement.

IV. METHODOLOGY

The system follows a structured pipeline comprising data preprocessing, feature selection, model training, and evaluation. Each stage is described below.

A. Data Preprocessing

Preprocessing is performed to ensure data quality and consistency before model training.

i) Column Removal: Columns with high uniqueness values (such as patient identifiers) are removed, as they do not contribute predictive information and may cause data leakage.

ii) Missing Value Imputation: Numerical columns with missing values are imputed using the column mean. Categorical columns are imputed using the most frequent value (mode). The SimpleImputer from Scikit-learn is employed for this purpose.

iii) Encoding: Categorical variables are converted to numerical format using one-hot encoding (pd.get_dummies), with the first category dropped to avoid multicollinearity.

B. Feature Selection

Feature selection is conducted using the SelectKBest method with the chi-square (χ^2) statistical test. The top 20 most significant features with respect to the target variable are retained. This reduces dimensionality, mitigates overfitting, and decreases computation time.

C. Feature Scaling

All selected features are standardized using StandardScaler, transforming each feature to have zero mean and unit variance. This is particularly important for Logistic Regression, which is sensitive to the scale of input features.

D. Train-Test Split

The preprocessed dataset is divided into an 80% training set and a 20% test set using a fixed random seed (random_state=42) to ensure reproducibility.

E. Classification Algorithms

i) Logistic Regression: A linear probabilistic classifier that models the log-odds of the target class as a linear combination of input features. Max iterations set to 1,000 for convergence.

ii) Decision Tree Classifier: A non-linear model that recursively partitions the feature space using splitting criteria such as Gini impurity or information gain. Prone to overfitting on training data.

iii) Random Forest Classifier: An ensemble of 100 decision trees trained on bootstrap samples of the data with random feature subsets at each split. Aggregates predictions via majority voting to reduce variance and overfitting.

V. RESULTS AND DISCUSSION

Each model was evaluated on the 20% holdout test set. Accuracy, precision, recall, F1-score, and confusion matrices were computed. An ROC curve was additionally plotted for the Random Forest model.

A. Model Performance Comparison

Table I: Classification Performance of ML Models

Algorithm	Accuracy	Precision	Recall
Logistic Regression	65.14%	0.64	0.65
Decision Tree	53.90%	0.52	0.54
Random Forest	60.69%	0.61	0.61

B. Analysis

Logistic Regression achieved the highest accuracy (65.14%) among the three models. Its performance can be attributed to the approximately linear separability of certain feature combinations in the dataset and its robustness against overfitting in high-dimensional spaces.

The Decision Tree model achieved the lowest accuracy (53.90%). This is consistent with known tendencies of decision trees to overfit training data, particularly when tree depth is unconstrained. The model demonstrated high variance, performing well on training data but generalizing poorly.

Random Forest achieved an intermediate accuracy of 60.69%. As an ensemble method, it reduced the variance of individual trees. The ROC curve analysis for Random Forest demonstrated its ability to discriminate between classes, though the AUC indicated moderate classification performance reflecting dataset complexity and class imbalance.

Overall, the results indicate that while all three models provide moderate predictive capability, none achieves the high accuracy (>80%) typically desired in clinical applications. This is partly attributed to class imbalance in the dataset and the inherent complexity of cardiovascular risk, which involves non-linear interactions among many variables.

VI. SYSTEM ARCHITECTURE

The system is structured as a modular pipeline:

- Data Input Module: Receives raw patient data from CSV or user interface
- Preprocessing Module: Handles missing values, encoding, and feature removal
- Feature Selection Module: Applies χ^2 test via SelectKBest
- Scaling Module: Normalizes features using StandardScaler
- Model Training & Testing Module: Trains and evaluates three classifiers
- Prediction Module: Accepts new patient input and returns risk classification
- Output Module: Displays result as “High Risk” or “Low Risk”

The modular design enables independent testing of each component and facilitates future extension, such as integration of additional algorithms or real-time data sources.

VII. CONCLUSION AND FUTURE WORK

This paper presented a machine learning-based heart attack risk prediction system trained on a dataset of 8,763 patient records. Three classifiers were implemented and compared. Logistic Regression outperformed Decision Tree and Random Forest with an accuracy of 65.14%, making it the most suitable model for this dataset. The system includes a real-time prediction function suitable for deployment in healthcare support applications.

Despite moderate accuracy levels, the system demonstrates the viability of machine learning as a decision-support tool in preventive cardiology. Future enhancements include:

- Addressing class imbalance using techniques such as SMOTE or cost-sensitive learning
- Exploring advanced algorithms including XGBoost, LightGBM, and deep neural networks
- Incorporating larger and more diverse real-world clinical datasets
- Integration with IoT-based wearable health monitoring devices for real-time risk assessment
- Developing a web or mobile interface for patient-facing deployment

With continued refinement, such systems hold significant promise for reducing cardiac mortality through timely intervention and preventive health promotion.

REFERENCES

[1] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, 2019.
[2] T. M. Mitchell, Machine Learning. McGraw-Hill Education, 1997.
[3] E. Alpaydin, Introduction to Machine Learning, 4th ed. MIT Press, 2020.
[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.
[5] Python Software Foundation, "Python Documentation," [Online]. Available: https://www.python.org
[6] Pandas Development Team, "Pandas Documentation," [Online]. Available: https://pandas.pydata.org
[7] NumPy Developers, "NumPy Documentation," [Online]. Available: https://numpy.org
[8] Scikit-learn Developers, "Scikit-learn Documentation," [Online]. Available: https://scikit-learn.org
[9] Matplotlib Development Team, "Matplotlib Documentation," [Online]. Available: https://matplotlib.org
[10] Kaggle, "Machine Learning Datasets and Tutorials," [Online]. Available: https://www.kaggle.com
[11] UCI Machine Learning Repository, "Heart Disease Dataset," [Online]. Available: https://archive.ics.uci.edu
[12] GeeksforGeeks, "Machine Learning Tutorials," [Online]. Available: https://www.geeksforgeeks.org