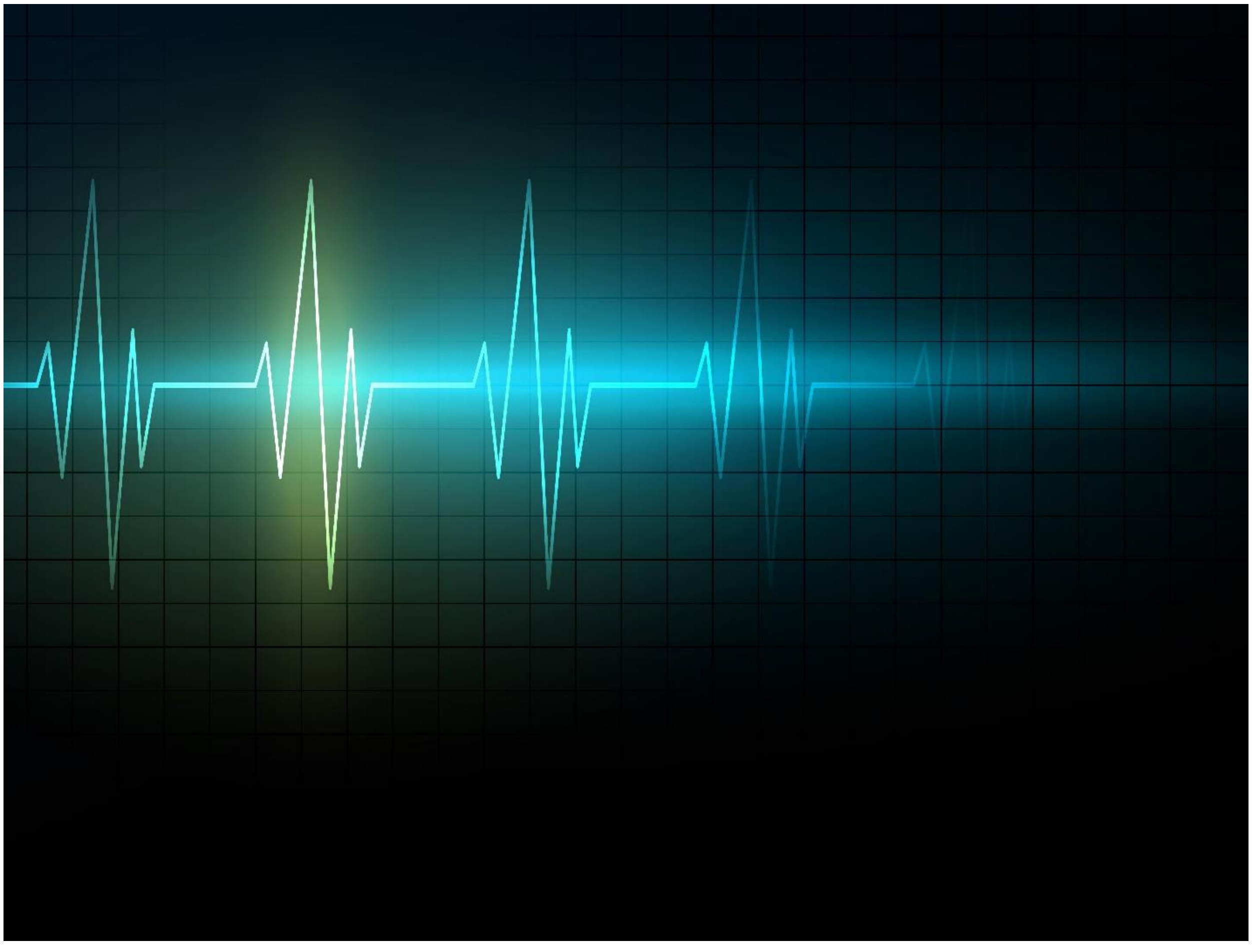
INTEGRATING STATISTICAL ANALYSIS AND SUPERVISED MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR DISEASE RISK PREDICTION: A COMPARATIVE EVALUATION.

BY: SAVIOUR DURU



designed by freepik

ABSTRACT

Cardiovascular disease (CVD) remains the leading global cause of mortality worldwide, highlighting the urgent need for early detection and prevention strategies. This research integrates traditional statistical analysis and supervised machine learning algorithms to predict heart disease risk using the Cleveland heart disease dataset from UCI machine learning repository comprising of 303 records with 13 clinical attributes.

Initial statistical techniques such as correlation analysis and hypothesis testing were employed to examine the relationships between variables such as age, chest pain type, resting blood pressure and the presence of heart disease. Results indicated that chest pain type, ST depression and the number of major blood vessels were significant predictors of heart disease risk aligning with medical knowledge.

Machine learning models including K-Nearest Neighbors (KNN), Logistic Regression, SVM, Random Forest, and Decision Tree were trained and among these, K-Nearest Neighbors achieved the highest performance with an accuracy of 91.8% and an ROC-AUC of 0.97. The results demonstrate that the combination of interpretable statistical tools and machine learning models can significantly enhance clinical decision making and preventive healthcare strategies.

1. INTRODUCTION

Cardiovascular disease is any disease or group of diseases affecting the heart and blood vessels, including coronary artery disease, stroke and heart failure. Despite advances in medicine, CVD remains a major health challenge. According to the 2021 Global Burden of Disease Study, over 19 million deaths were attributed to CVD worldwide. The growing volume of healthcare data presents opportunities for data driven-prediction and prevention.

Machine learning (ML) has emerged as a revolutionary approach in healthcare, capable of identifying hidden patterns in clinical data. However, the interpretability of ML models is often limited. This study bridges the gap between statistical interpretability and predictive accuracy but combining classical statistical methods with modern machine learning algorithms.

2. LITERATURE REVIEW

2.1 Machine Learning – Historical Foundations

The theoretical and historical foundations of machine learning date back to early models of neural activity and formal work on a machine intelligence. Seminal contributions include McCulloch & Pitts' formulation of a logical calculus for nervous activity (which inspired later neural network research), Alan Turing's 1950 articulation of machine intelligence, and Arthur Samuel's early empirical work on machine learning with checkers in 1959. These milestones established key ideas, artificial neurons, algorithmic learning, and the experimental study of learning machine that underpin modern ML tools such as scikit-learn. *(McCulloch & Pitts, 1943); Turing (1950); Samuel (1959).*

2.2 Machine Learning in Healthcare

Applied to medicine, machine learning has been shown to augment diagnostic accuracy and image interpretation by extracting patterns from large, heterogeneous clinical datasets. Review and studies emphasize ML's potential in cardiac imaging and diagnostic support enabling faster interpretation of scans and aiding clinicians with decision support while also highlighting the need for rigorous validation before clinical deployment. Large-scale evaluations indicate ML can improve risk stratification when trained and validated on routing clinical data. *(Dilsizian & Siegel, 2014).*

2.3 Machine Learning Application in Cardiovascular Research

In cardiovascular research specifically, a broad set of method from support vector machines and decision trees to neural networks and convolutional networks (CNNs) have been applied for diagnosis, risk prediction, and image analysis. Studies report successful CAD (coronary artery disease) detection from HRV/ECG signals using SVM and related methods, while CNNs and other deep architectures have advanced automated interpretation of cardiac imaging and ECGs, improving feature extraction and diagnostic throughput. Ensemble methods and feature-selection approaches further boost predictive performance and interpretability in heart-disease models. *(Dolatabadi et al., 2017); (Systematic reviews of deep learning in cardiac imaging, JACC review, 2019)*

3. METHODOLOGY

3.1 Machine Learning – Historical Foundations

The Cleveland Heart Disease dataset (UCI Machine Learning Repository) contains 303 patient records with 14 attributes including age, gender, cholesterol, resting blood pressure, and exercise - induced angina. The target variable indicated the presence of heart disease (1) or absence of heart disease (0).

3.2 Data Preprocessing

- **Handling Missing Values:** Records with missing values were identified and removed to ensure data quality
- **Encoding:** Categorical variables were transformed into numerical representations e.g. chest pain type, thalassemia.
- **Scaling:** Continuous variables were standardized e.g. cholesterol.
- **Split:** 80% training and 20% testing using stratified sampling.

3.3 Statistical Analysis

Table 3.1: Statistical tests and formulas

Test	**Formula**
Pearson’s correlation analysis	$r = \frac{\sum(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{(x_i - \overline{x})^2(y_i - \overline{y})^2}}$
Independent sample T-test	$\boldsymbol{t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$
Chi-square test	$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Analysis of Variance (ANOVA)	$F = \frac{Between\ Groups}{Wthin\ Groups}$

Descriptive statistics summarized the population characteristics. Correlation analysis (Pearson’s r) identified relationships such as:

- Age negatively correlated with maximum heart rate (r = -0.39).
- Age positively correlated with blood pressure (r = 0.28).

Hypothesis tests (t-test, ANOVA, Chi-square) showed:

- Age and sex significantly associated with heart disease ($p < 0.05$).
- Chest pain type strongly predicted presence of heart disease.

3.4 Model Development

Five algorithms were implemented using Python (Scikit-learn). Hyperparameter tunning was performed using GridSearchCV with 5-fold cross-validation.

Table 3.2: Machine learning models and description

Model	Description
Logistic Regression	Simplicity, interpretability and ability to model linear relationships.
Decision Tree	Capturing non-linear relationships and easy visual interpretation.
Random Forest	Improved performance and reduced overfitting through ensemble learning.
SVM	Robust in high dimensional spaces
KNN	Instance based approach and non-parametric nature

3.5 Model Evaluation Metrics

To ensure consistent performance assessment across models, standard evaluation metrics were computed. These include Accuracy, Precision, Recall (Sensitivity), and F1-Score. These metrics captures the correctness, completeness and robustness of the predictions, the formulas are shown in Table 3.3

Table 3.3: Evaluation metrics and formulas

Metric	Formula	Description
Accuracy	Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$	Proportion of total correct predictions
Precision	Precision = $\frac{TP}{TP + FP}$	Measures how many predicted positives are actual positives
Recall (Sensitivity)	Recall = $\frac{TP}{FP + FN}$	Measures how well the model detects actual positives.
F1-Score	F1-Score = $\frac{2 \times Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall
ROC-AUC	Area under the ROC curve	Reflects the model's ability to discriminate between positive and negative classes.

Confusion Matrix: This is a simple table used to measure how well a classification model is performing. It compares the predictions made by the model with the actual results and shows where the model was right or wrong and it is broken down into 4 categories.

- True Positive (TP): The model correctly predicted a positive outcome i.e. the actual outcome was positive.
- The Negative (TN): The model correctly predicted a negative outcome i.e. the actual outcome was negative.
- False Positive (FP): The model correctly predicted a negative outcome i.e. the actual outcome was negative. It is also known as Type 1 error.
- False Negative (FN): The model incorrectly predicted a negative outcome i.e. the actual outcome was positive. It is also known as Type II error

It helps calculate key measures like accuracy, precision and recall which gives a better idea of performance especially when the data is imbalanced.

4. RESULTS AND DISCUSSION

4.1 Overview of Data and Statistical findings

Descriptive analysis showed an average patient age of 54.4 years (SD = 9.0) and mean resting blood pressure of 131.7mmHg. Cholesterol levels averaged 246.7mg/dl, with most patients exhibiting moderate hypercholesterolemia.

Correlation results revealed a negative relationship between age and maximum heart rate (r = -0.39), consistent with the physiological decline of cardiac capacity with age. Chi-square and t-tests further indicated significant associations between sex, chest pain type and presence of heart disease ($p < 0.05$), confirming their predictive importance.

Notably, chest pain type 4 (asymptomatic) occurred predominantly in patients with diagnosed heart disease, highlighting the diagnostic relevance of symptomless ischemic cases

4.2 Model Performance Comparison

Table 4.1: Summary of model performance

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
KNN	0.9180	0.8965	0.9286	0.9122	0.9681
Logistic Regression	0.8688	0.8125	0.9286	0.8666	0.9599
SVM	0.8524	0.7878	0.9286	0.8524	0.9502
Random Forest	0.8852	0.8387	0.9286	0.8813	0.9426
Decision Tree	0.8688	0.8571	0.8571	0.8571	0.8674

KNN achieved the **highest performance** (Accuracy = 91.8%, AUC = 0.97), demonstrating its string discriminative power in classifying patients correctly.

Logistic Regression and Random Forest slightly trailed KNN, they offered better interpretability, making them valuable for clinical deployment.

4.3 Model Interpretability and Feature Importance

Feature importance across models revealed consistent predictors of heart disease. The top features are:

- Ca (number of major vessels)
- Old peak (ST depression by exercise)
- Cp (chest pain type)
- Thal (thallium stress test results)
- Thalach (maximum heart rate achieved)

Random Forest and Logistic Regression analyses both confirmed that **higher ST depression** and **more obstructed vessels** significantly increased the likelihood of heart disease.

Conversely, higher maximum heart rate and female gender were associated with **lower disease risk**, supporting known clinical trends.

4.4 Discussion

The results align with prior cardiovascular research, which identifies **ST depression, vessel obstruction** and **chest pain type** as critical diagnostic features.

The high performance of KNN suggests that distance-based learning can effectively capture nonlinear relationships in small, structured medical datasets.

For real world applications, Logistic Regressions remains advantageous because of its interpretability, clinicians can trace how each variable contributes to disease risk.

Random Forest, while slightly less accurate, offers robustness and useful feature ranking, making it a strong candidate for clinical decision support systems.

5. CONCLUSION

This study demonstrated that machine learning models, supported by statistical validation can accurately predict cardiovascular disease using routine clinical features.

Among the five algorithms tested, K-Nearest Neighbors (KNN) achieved the best overall performance (Accuracy = 91.8%, AUC = 0.97).

Key predictors include number of major vessels (ca), ST depression (oldpeak), chest pain type, and thallium test results.

The integration of both traditional statistical methods (for interpretability) and machine learning algorithms (for predictive precision) ensures findings are not only predictive but also explainable which yields a balanced approach for medical analytics.

Such hybrid frameworks can improve early diagnosis, risk assessment, and resource prioritization in healthcare system

Practical implications suggests that hospitals and clinics can integrate these predictive models into clinical decision support systems (CDSS) for early screening.

Also, non-invasive data such as ECG readings, chest pain patterns, and exercise tolerance can serve as preliminary risk indicators in community health drives.

Lastly, Public health interventions can prioritize early testing for individuals exhibiting significant ST depression or multiple blocked vessels.

6. REFERENCES

Ahmed, Intisar, "A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING AND DATA MINING" (2022). Electronic Theses, Projects, and Dissertations. 1591.

American Heart Association. (2023). *Heart Disease and Stroke Statistics 2023 Update: A Report From the American Heart Association*. Circulation.

Davari Dolatabadi A, Khadem SEZ, Asl BM. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. Comput Methods Programs Biomed. 2017 Jan; 138:117-126. doi: 10.1016/j.cmpb.2016.10.011. Epub 2016 Oct 24. PMID: 27886710.

Dilsizian, S.E., Siegel, E.L. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. *Curr Cardiol Rep* **16**, 441 (2014).

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of Scikit-learn package in Python programming language. *Journal of Educational and Behavioral Statistics, 44*(3), 348–361.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository.

McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943).

Pandey, S., & Singh, R. (2018). A review on heart disease prediction using machine learning techniques. *Procedia Computer Science, 132*, 937–944.

Udomboso, C. G., Sigauke, C., & Adinya I., (2025). *"Fusion Sampling Validation in Data Partitioning for Machine Learning,"* arXiv preprint arXiv:2508.01325.

Vabalas A. E., Gowen, Poliakoff E., and Casson A. J., *"Machine learning algorithm validation with a limited sample size,"* PLoS One, vol. 14, no. 11, pp. 1–20

Weng, S., Kai, J., Guo, Y., Qiao, Q., & Ljung, T. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE, 12*(4), e0174944.