International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 09 September-2026 | Impact Factor: 3.5

A Comparative Analysis of Machine Learning Algorithms for Heart Diseases Prediction

Hardik Varma[1], Aryan M. Sinha[2]

[1] Information Technology / Bharati Vidyapeeth University College of Engineering, Pune, India
[2] Information Technology / Bharati Vidyapeeth University College of Engineering, Pune, India

Corresponding Author Email: sinhaaryan399@gmail.com

How to Cite this Article:

Varma, H. & Sinha, A. M. (2026). A Comparative Analysis of Machine Learning Algorithms for Heart Diseases Prediction. International Journal of Creative and Open Research in Engineering and Management, 02(9), 1-9. https://doi.org/10.55041/ijcope.v2i8.302

https://doi.org/10.55041/ijcope.v2i8.302

Abstract—

The term heart disease refers to the range of conditions that impacts heart and its blood vessels in a negative manner. This has led to increase mortality rate worldwide and a major health issue in recent years and continues to have increasing statistics in future. Modern day lifestyles and choices have made it increasingly common, not only in aged population but also in younger adults. New generation advanced medicines are fortunately a breakthrough in assisting and curing the disease, but conventional techniques are only optimal to provide on the spot single data reading and result, it fails to predict the future circumstances and possible emergency which may be caused due to underlying issue, or provide with early diagnosis to facilitate prompt medical assistance. This is where Machine learning (ML) comes into the picture, delivering the required and necessary early analysis of cardiovascular disease (CVD) through validating various features among thousands of data in a dataset. This study has developed predictive models and used Cleveland dataset with 11 different features namely age, sex, cholesterol, resting heart rate, exercise-induced angina and other health-related indicators that make up this dataset. There are four different predictive models that we will be using here to predict the health analysis of heart related disease, which are Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbor (KNN). All of the mentioned models are developed and used under Google colab platform. There are 4 different datasets consisting of Cleveland, Hungary, Switzerland and VA Long Beach which are preprocessed and cleaned and then being used and tested here with machine learning models to test its accuracy and other result aspects. The best result and performance among these four models tested across each of the datasets and having minimum spread (max-min) is of Support Vector Machine (SVM). with an F1 score of 0.7732. This study will focus on developing multi predictive models to help pave a way to ensure early diagnosis and provide early treatment.

Keywords— heart; disease; predictive; machine learning; models

I. INTRODUCTION

Heart, one of the major organs in human body, consists of two upper atria (receiving chambers), two lower ventricles (pumping chambers), septum which separates the two sides of heart, valves and arteries. Additionally, the overall function primarily starts with the heart walls and electrical system of heart. The outer layer (epicardium), thick muscular layer (myocardium) responsible for contractions and the inner endocardium. It's rhythm is controlled by the electrical conduction system, which coordinates the heart's muscles to contract and pump blood at a normal paced rhythm allowing the blood to flow into the body and vice versa.

Any restriction, blockage, electrical misconduction or irregularities in heart can cause a serious medical emergency.

Cardiovascular disease (CVD) has been the leading cause of death globally, with increasing statistics every year with 18.5 million in 2023, 20.5 million in 2025 and projected to 35.6 million by 2050 by European Journal of Preventive Cardiology. Blockage in artery due to fat or calcium deposits, congenital disease, cardiomyopathies, hypertension are the upmost reasons.

Further, unhealthy or sedentary lifestyle, smoking, substance abuse, poor diet, bad stress management also increases the risk of occupying these diseases.

Conventional techniques are accurate, reliable and are among the options for treating and evaluating patients, such as angiography, echocardiogram (ECG), stent insertion, statins, vasodilators and many more. But these are invasive procedures, resource-intensive and have varied effectiveness among different people.

Technological advancements have significantly impacted the prevention and treatments in a positive manner. Techniques such as gene editing (One-time in vivo CRISPR editing of PCSK9/ANGPTL3; potential "one-and-done" lipid therapy), injections such as Inclisiran (siRNA PCSK9i, sustained LDL-C reduction ~50%), SGLT2 inhibitors (Cardioprotective beyond glycaemic control), Lp(a)-targeting RNA therapies (80–95% Lp(a) reduction; targeting a previously "untreatable" genetic risk factor) have promising solutions in reducing the major risk factors of CVD, by vast difference.

One of the important areas of curing and preventing CVD is early diagnosis, detection and accurate future health marker prediction, which somehow cannot be tackled using conventional techniques, as most of the patients or people are reluctant to participate in clinical trials, routine health checkups. Here Machine learning (ML) plays a big and important part, defining a new way to not only identify and diagnose CVD, but with low-cost and reasonable accuracy. ML techniques for diagnosis does not require multiple clinical trials, as different records of various patients with their features can be used ethically to predict and evaluate CVD with high accuracy.

It is expected that ML application will be globally used to increase doctor's work efficiency, generate economic benefits.

II. LITERATURE REVIEW

Machine learning has become a central methodology in cardiovascular disease (CVD) prediction research, with studies varying in dataset choice, feature selection strategy, and algorithmic approach.

- Dayana and Nandini (2024) conducted a comparative analysis using a large dataset (13,861 test instances) incorporating features such as age, gender, cholesterol levels, blood pressure, and electrocardiogram results, selected through a structured preprocessing pipeline involving data cleaning, normalization, and domain-informed feature selection. They evaluated Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, and XGBoost, finding that Random Forest, Gradient Boosting, and XGBoost each achieved the highest accuracy (0.74), attributing this to the strength of ensemble techniques in combining multiple learners to produce more robust predictions, while KNN performed weakest (0.68) due to its sensitivity to complex decision boundaries. Similarly, Anjum et al. (2024) used a manually collected dataset of 600 heart failure patients with 13 clinical attributes (including ejection fraction, serum creatinine, and serum sodium) drawn from Bangladeshi hospitals, applying Logistic Regression, SVM, XGBoost, LightGBM, Decision Tree, and Bagging. Their results identified XGBoost as the top performer (94.80% accuracy, 90.0% AUC), followed closely by LightGBM (92.50% accuracy, 92.00% AUC), reinforcing the recurring finding that boosting-based ensemble methods outperform simpler classifiers in CVD prediction tasks.

- Ogunpola et al. (2024) extended this line of inquiry by explicitly addressing the challenge of imbalanced datasets, a limitation they noted was often overlooked in prior work. Using two datasets—the Cardiovascular Heart Disease Dataset (1,000 samples, 13 attributes) from Mendeley and the Heart Disease Cleveland Dataset (303 patients, 14 attributes) from Kaggle—they applied preprocessing techniques including oversampling, feature scaling, and dimensionality reduction, and tested seven classifiers:

KNN, SVM, Logistic Regression, CNN, Gradient Boosting, XGBoost, and Random Forest. After hyperparameter tuning via GridSearchCV, XGBoost achieved the best results on the Cardiovascular Heart Disease Dataset (98.50% accuracy, 99.14% precision, 98.29% recall, 98.71% F1-score), while KNN with k=7 performed best on the Cleveland dataset (91.80% accuracy), leading the authors to conclude that both KNN and XGBoost consistently ranked as top performers across datasets. This dual-dataset design highlights how model performance can vary depending on dataset characteristics, even when the same algorithms are applied.

Complementing these empirical studies,

- Rao and Muneeswari (2024) provided a broader review synthesizing findings from 27 studies published between 2021 and 2023, categorizing approaches into machine learning-based, ensemble-based, data mining-based, IoT-based, and deep learning-based prediction methods. Their synthesis reported that models such as HyOPTXg (Srinivas & Katarya, cited in Rao & Muneeswari, 2024) achieved 94.7% accuracy using Optuna-tuned XGBoost, while ensemble frameworks combining Random Forest, Naïve Bayes, and Gradient Boosting—selected via feature evaluators such as Gain Ratio and Information Gain—generally produced higher detection accuracy than single-classifier approaches. The review also noted that Bi-LSTM approaches using IoT-derived data achieved accuracy as high as 98.86%, though such methods incurred greater computational cost and training complexity. Notably, Rao and Muneeswari (2024) concluded that despite the diversity of methodologies reviewed, most approaches suffered from long computation times and comparatively low accuracy attributable to the inclusion of irrelevant attributes, indicating that feature selection remains an unresolved challenge across the field.

Across these studies, a consistent trend emerges: ensemble and boosting algorithms—particularly XGBoost, Random Forest, and Gradient Boosting—repeatedly outperform simpler linear or instance-based models such as Logistic Regression and KNN, largely due to their capacity to combine weak learners and correct sequential errors. However, notable differences exist in dataset scale (ranging from 303 to over 13,000 instances), feature composition, and preprocessing rigor, which likely account for the wide variance in reported accuracy (from roughly 68% to over 98%) across studies. Furthermore, while several works highlight the importance of addressing imbalanced data and irrelevant feature inclusion, only a subset of studies (e.g., Ogunpola et al., 2024) explicitly implement oversampling or systematic feature selection frameworks. This inconsistency in methodological rigor, combined with the review's identified gap concerning long computation times and insufficient generalizability across diverse datasets, underscores the need for a comparative framework—such as the one proposed in the present study—that systematically evaluates multiple algorithms under standardized preprocessing conditions to establish more reliable and reproducible benchmarks for CVD detection.

III. METHODOLOGY

The methodology followed in this work encompasses data collection, preprocessing, experimental setup, model selection, and performance evaluation. A complete description of each stage is provided below.

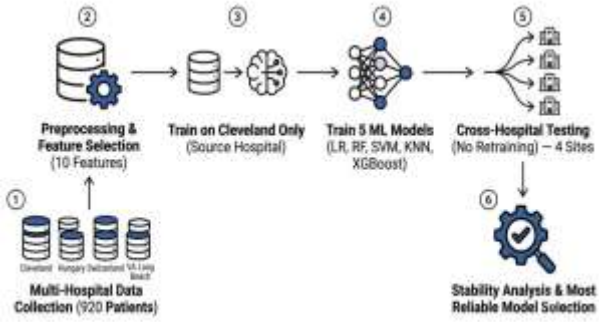


Figure 1: Cross-Hospital Generalization Architecture for Heart Disease Prediction

Figure 1 illustrates the complete workflow: multi-hospital data collection from four sites (Cleveland, Hungary, Switzerland, VA Long Beach), preprocessing and feature selection (10 common features), training exclusively on Cleveland (source hospital), training five machine learning models, cross-hospital testing on all four sites without retraining, and final stability analysis to identify the most reliable model.

III(I). Dataset Description For our experiments:

We used the well-established UCI Heart Disease repository. This dataset contains records from 920 patients drawn from four independent medical institutions: Cleveland, Hungary, Switzerland, and the VA Long Beach center. The dataset is publicly available and has been extensively utilized in predictive modeling studies.

We narrowed down the feature set to 10 variables that were consistently recorded across all four sites. These ten features are: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression (oldpeak).

We simplified the original multiclass target into a binary label—0 for no disease and 1 for disease present. Several features were excluded from the analysis because they were largely missing in the Hungary, Switzerland, and VA Long Beach datasets, and imputing them would have introduced more guesswork than signal.

III(II). Data Preparation and Split Strategy:

All data cleaning and preprocessing steps were performed using only the Cleveland partition. Missing numeric values were replaced with the median from Cleveland, and missing categorical values were replaced with the mode from Cleveland. This decision was critical to prevent information from the other three hospitals from influencing the training data. For scaling, we applied standard normalization (zero mean, unit variance). The normalization parameters were learned exclusively from the Cleveland training portion and then applied to all other datasets. This simulates a realistic scenario in which a model is developed in one site and then transported to others. The training set was formed by randomly selecting 80% of the Cleveland records. The remaining 20% of Cleveland, along with the complete datasets from Hungary, Switzerland, and VA Long Beach, were set aside as evaluation partitions.

III(III). Novelty of the Experimental Design:

The key differentiator of our study lies in the evaluation scheme. Most earlier works have combined all sites. and performed random splits, we intentionally trained classifiers on a single site. This allows us to measure population-level generalizability, which is a more demanding and practically relevant test than within-sample accuracy. Our focus is not simply on predictive power, but on consistency across varying clinical populations.

III(IV). Model Selection We selected a set of five models, each representing different learning paradigm, to ensure a comprehensive comparison of model behavior across hospital populations.

The chosen models include:

• Logistic Regression (linear classifier)

• Random Forest (bagged ensemble of trees)

• Support Vector Machine (SVM)

• K-Nearest Neighbors (distance-based lazy learner)

• XGBoost (boosted tree model)

All models were implemented using their default hyperparameters, with a fixed random seed (random_state = 42) applied throughout to ensure reproducibility. The one exception was XGBoost's learning rate, which was reduced from its default value of 0.3 to 0.1.

III(V). Evaluation Protocol:

For evaluation of all models a range of standard classification metrics were computed to provide a comprehensive assessment of predictive ability of the models. This included:

• Accuracy – the proportion of correctly classified instances out of the total predictions. • Precision – the ratio of true positive predictions to the total positive predictions, indicating how often the model is correct when it predicts disease.

• Recall (Sensitivity) – the ratio of true positive predictions to the actual positive instances, reflecting the model's ability to correctly identify patients with heart disease.

• F1 Score – the harmonic mean of precision and recall. It provides a balanced measure of performance, making it especially useful when both false positives and false negatives carry significant consequences, as is the case in medical prediction tasks.

The F1 Score was prioritized due to its balance between false alarms and missed detections.

To assess how consistently each model performs across different hospital populations, we introduced a stability metric called "Spread." Spread is defined as the range between the maximum and minimum F1 Score achieved by a model across the four hospitals.

Lower spread values indicate more consistent performance across heterogeneous populations, which we interpret as a proxy for methodological robustness rather than a claim of clinical readiness.

Table 1 presents the complete set of results—Accuracy, Precision, Recall, and F1 Score— for each of the five models across all four hospitals. The Cleveland column represents performance on the held-out 20% test split from the training hospital, while the remaining columns reflect performance on entirely unseen patient populations.

Model	Hospital	Accuracy	Precision	Recall	F1 Score
Logistic Regression	Cleveland	0.787	0.808	0.724	0.764
Logistic Regression	Hungary	0.823	0.741	0.783	0.761
Logistic Regression	Switzerland	0.423		0.391	0.559
Logistic Regression	VA Long Beach	0.695	0.824	0.752	0.786
Random Forest	Cleveland	0.787	0.786	0.759	0.772
Random Forest	Hungary	0.819	0.762	0.726	0.744
Random Forest	Switzerland	0.642	0.961	0.643	0.771
Random Forest	VA Long Beach	0.710	0.832	0.765	0.797
SVM	Cleveland	0.787	0.808	0.724	0.764
SVM	Hungary	0.819	0.752	0.752	0.749
SVM	Switzerland	0.659	0.974	0.652	0.781
SVM	VA Long Beach	0.710	0.827	0.772	0.799
KNN	Cleveland	0.787	0.808	0.724	0.764
KNN	Hungary	0.788	0.729	0.660	0.693
KNN	Switzerland	0.593	0.971	0.583	0.728
KNN	VA Long Beach	0.685	0.831	0.725	0.774

V. RESULTS AND DISCUSSION

This section presents the experimental findings obtained from evaluating five machine learning models across four hospital populations. All models were trained exclusively on the Cleveland dataset and tested on the remaining three hospitals without any retraining. We report standard classification metrics, with particular emphasis on F1 Score and the stability metric (Spread) introduced in the previous section.

Overall Performance

Observations:

Several patterns emerge from Table 1. First, all models achieved reasonably strong performance on the Cleveland test set, with F1 scores ranging from 0.746 to 0.772. This confirms that the models were able to learn meaningful patterns from the training data. Second, performance on the Hungary and VA Long Beach datasets remained comparable to Cleveland for most models, with F1 scores ranging from 0.693 to 0.814. This suggests that these hospital populations share similar underlying data distributions with Cleveland. Third, a notable drop in performance was observed on the Switzerland dataset across all models. SVM (0.781) and Random Forest (0.771) maintained the highest F1 scores on Switzerland, while KNN (0.728), and particularly Logistic Regression (0.559) showed more substantial degradation.

Fourth, Logistic Regression exhibited a striking pattern on Switzerland: while its precision remained very high (0.978), its recall collapsed to 0.391. This indicates that the model became extremely conservative on this population, correctly identifying only 39% of actual positive cases.

Cross-Hospital Stability Analysis:

To quantify how consistently each model performed across different hospital populations, we computed the Spread—defined as the difference between the highest and lowest F1 Score achieved by a model across the four hospitals. Lower spread values indicate more stable performance across heterogeneous populations.

Table 2: F1 Scores and Spread Across Hospitals

Model	Cleveland	Hungary	Switzerland	VA Long Beach	Spread (Max-Min)
SVM	0.764	0.749	0.781	0.799	**0.050**
Random Forest	0.772	0.744	0.771	0.797	**0.053**
KNN	0.764	0.693	0.728	0.774	0.081
XGBoost	0.746	0.712	0.724	0.814	0.102
Logistic Regression	0.764	0.761	0.559	0.786	0.227

Key Findings: SVM achieved the lowest spread (0.050), making it the most stable model across all four hospital populations. Its F1 scores remained tightly clustered between 0.749 and 0.799. Random Forest was a close second with a spread of 0.053, demonstrating that tree-based bagging ensembles also exhibit strong cross-hospital generalizability. KNN (spread = 0.081) and XGBoost (spread = 0.102) showed moderate stability. Logistic Regression exhibited the largest spread by a substantial margin (0.227), driven almost entirely by its collapse on the Switzerland dataset (F1 = 0.559). This highlights the vulnerability of linear models to population-level distribution shifts.

Analysis of Logistic Regression Collapse on Switzerland

A closer examination of Logistic Regression's performance on Switzerland reveals a critical diagnostic pattern: Precision: 0.978 — extremely high, meaning that when the model predicted disease, it was almost always correct. Recall: 0.391 — extremely low, meaning the model failed to identify 61% of actual positive cases.

This pattern indicates that the model became highly conservative on the Switzerland population. It only predicted the positive class when the evidence was overwhelmingly strong, effectively prioritizing precision at the expense of recall. This behaviour is consistent with a linear model that relies on feature relationships that do not generalize across populations with different underlying distributions.

A likely contributing factor is the near-zero variance in serum cholesterol (chol) within the Switzerland subset. Since Logistic Regression depends on meaningful linear relationships, the lack of variation in this feature would have prevented the model from establishing a useful coefficient, degrading its ability to discriminate effectively on this population.

In contrast, SVM and Random Forest maintained strong recall on Switzerland (0.652 and 0.643 respectively), demonstrating their robustness to such distributional shifts. This underscores the advantage of non-parametric and ensemble methods when deploying models across heterogeneous clinical settings.

Visual Comparison of Stability

To provide a clear visual representation of these findings, we generated two figures. Figure 2 shows the F1 Score trajectories for each model across the four hospitals. The lines for SVM and Random Forest remain relatively flat, visually confirming their stability. Logistic Regression's line exhibits a pronounced dip at Switzerland, highlighting its vulnerability to population shifts.

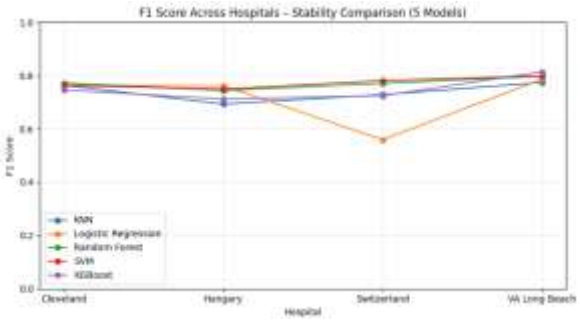
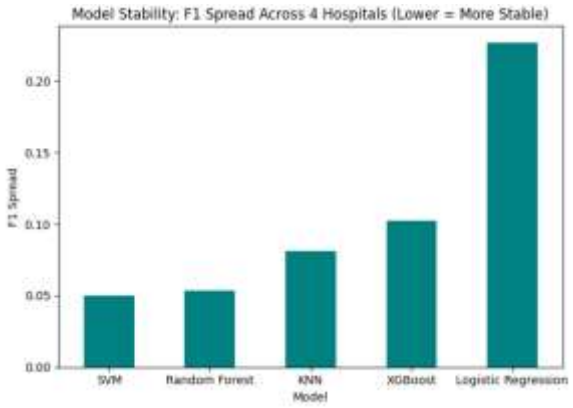


Figure 3 presents the Spread values as a bar chart. Lower bars indicate more stable models. SVM and Random Forest have the shortest bars, while Logistic Regression's bar is substantially taller, reinforcing its status as the least reliable model for cross-hospital deployment.



V. CONCLUSION

The experimental findings can be summarized as follows: Most Stable Models: SVM (Spread = 0.050) and Random Forest (Spread = 0.053) demonstrated the highest cross-hospital consistency. These models are recommended for deployment scenarios where patient populations may vary significantly. Least Stable Model: Logistic Regression (Spread = 0.227) failed to generalize to the Switzerland population due to extremely low recall (0.391). This highlights the limitations of linear models in cross-population settings. Moderate Performers: KNN (Spread = 0.081). While KNN maintained reasonable consistency. Switzerland as a Challenging Population: All models performed worst on Switzerland, indicating that this hospital's patient demographics or data quality differ substantially from the others. However, SVM and Random Forest absorbed this shift better than the remaining models.

REFERENCES

[1] arXiv:2405.17059 **[cs.LG]**

arXiv:2405.17059v1 **[cs.LG]**

https://doi.org/10.48550/arXiv.2405.17059

[2]

https://www.researchgate.net/publication/393747946_Deep_Learning_for_Cardiovascular_Disease_Detection_A_Review_Based_on_Cardiac_Magnetic_Resonance_Imaging_Data

[3]
https://www.researchgate.net/publication/380924765_ELM_and_LIGHTGBM_A_Hybrid_Machine_Learning_Technique_with_Intelligent_IOT_to_Predict_the_cardiovascular_disease