

Machine Learning Model for Heart Disease Detection: A Comparative Analysis of SVM vs KNN

Uttam Kumar Giri

Chitkara University School of Engineering
and Technology, Chitkara University,
Himachal Pradesh, India
uttameast@gmail.comline

Merry Saxena

Department of Interdisciplinary Courses in
Engineering, Chitkara University Institute of
Engineering & Technology, Chitkara
University, Punjab, India
merry.saxena@chitkara.edu.in

Dibyahash Bordoloi

Dept of CSE, Graphic Era Hill University
Graphic Era Deemed to be University
Dehradun, India
dibyahash.cse@geu.ac.in

Ramamani Tripathy

Chitkara University School of Engineering
and Technology, Chitkara University,
Himachal Pradesh, India
ramamani.tripathy@chitkarauniversity.edu.in

Srikanta Kumar Mohapatra

Chitkara University Institute of Engineering
and Technology,
Chitkara University, Punjab, India
srikanta.2k7@gmail.com

Pradeepta Kumar Sarangi

Chitkara University School of Engineering
and Technology, Chitkara University
Himachal Pradesh, India
pradeeptasarangi@gmail.com

Abstract—Abstract: Out of many critical medical issues, heart disease is one that primarily impacts the cardiovascular system, including the heart and its associated organs. Individuals who smoke, suffer from high blood pressure, have elevated cholesterol levels, maintain poor dietary habits, lack physical activity, or are overweight or obese face a greater risk of developing heart disease. The most frequently diagnosed cardiovascular condition is coronary artery disease (CAD), although other forms of heart disease include congestive heart failure, arrhythmias, and congenital heart defects. This condition is often referred to as cardiovascular disease. Timely and precise diagnosis is essential for saving lives and preventing further complications. Recent advancements in machine learning diagnostic techniques have demonstrated enhanced reliability and effectiveness in detecting heart disease. In this work, we have implemented a SVM algorithm followed by a KNN algorithm to accurately identify the early stages of heart disease within a specific dataset. The algorithms consider various factors, including age, gender, resting blood pressure, cholesterol levels, blood sugar levels, and ECG results when making predictions. The experimental findings reveal that the SVM model outperforms the others, achieving an impressive accuracy rate of 89% on the UCI dataset.

Key words:

Keywords—*heart disease detection, nearest neighbor, support vector, machine learning*

I. INTRODUCTION

Cardiovascular disease is widely recognized as a serious chronic health issue worldwide. Often, a cardiac condition is not diagnosed until a person exhibits signs of heart failure, arrhythmia, or a heart attack. A heart attack is typically marked by pain or discomfort in the chest, along with symptoms like nausea or pain in the arms. Other potential indicators of heart disease include heartburn, severe fatigue, and pain in the upper body or back. Heart failure presents with various symptoms, such as swelling in the feet, abdomen, or neck veins. Arrhythmia, another type of cardiovascular disorder, is characterized by palpitations in the chest. The World Health Organization (WHO) reports that heart disease

is the foremost cause of death worldwide. [1]. The development of non-invasive diagnostic methods has not only decreased mortality rates associated with heart disease but has also significantly reduced the risks linked to invasive diagnostic procedures [2].

Machine learning has demonstrated its effectiveness in a range of medical applications, including predicting Covid [3] and diagnosing plant leaf diseases. This study aims to develop and evaluate a comparable method for detecting heart disease through the use of KNN and SVM algorithms.

II. LITERATURE REVIEW

Many researchers have utilized a range of machine learning algorithms, such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees (DT), in their efforts to create non-invasive healthcare systems as alternatives to invasive procedures, with the goal of reducing related complications [4]. Using different data sets and algorithms, researchers have reported different results [5]. Table- 1 highlights some notable contributions by researchers to the related work discussed here.

TABLE I SIMILAR WORKS IN THIS FIELD

Reference	Dataset Used	Algorithm Used	Best Accuracy
[6]	Kaggle	Multiple algorithms	87.28%
[7]	UCI Dataset	Combination of RF with a Linear model	88.7%
[8]	UCI Dataset	RF	88.35%
[9]	CVD and Framingham	Multiple classifiers	75%
[10]	UCI Dataset	Logistic Regression	82.89%
[11]	UCI Dataset	NB, DT	82.7%
[12]	UCI Dataset	ANN, SVM	87.5%
[13]	Cleveland heart disease	Neural Network Ensemble method	89.01%

The three key components of a dataset, classifier, and accuracy can be elaborated as follows:

In some case the machine learning parameters are used efficiently to showcase the observation [14-16].

a. ****Dataset**** - Currently, there is no standardized dataset that meets the requirements for this purpose. The UCI Dataset is often preferred by researchers due to its accessibility and thoroughness.

b. ****Classifiers**** - A variety of classifiers have been proposed as the best options for implementation in this context. SVM and KNN are two of the most popular algorithms.

c. ****Accuracy**** - It has been noted that accuracy is not stable; instead, it fluctuates depending on the dataset and classifier used.

III. OBJECTIVE

This research seeks to develop two distinct machine learning models, specifically SVM and KNN, to predict heart disease utilizing the publicly accessible UCI dataset.

IV. DATA SET

The dataset used to implement the machine learning models are retrieved from the official website of the University of California, Irvine [17]. Figure 1 provides a description of the sample data set, which consists of 493 input rows and 14 attributes.

exang	oldpeak	slope	ca	thal	target
0	0.0	2	0	2	1
0	0.0	2	0	2	1
0	0.0	2	0	2	1
0	0.0	2	0	2	1

Fig. 1: Sample dataset

Figure 2 illustrates the distribution of patients with and without heart disease.

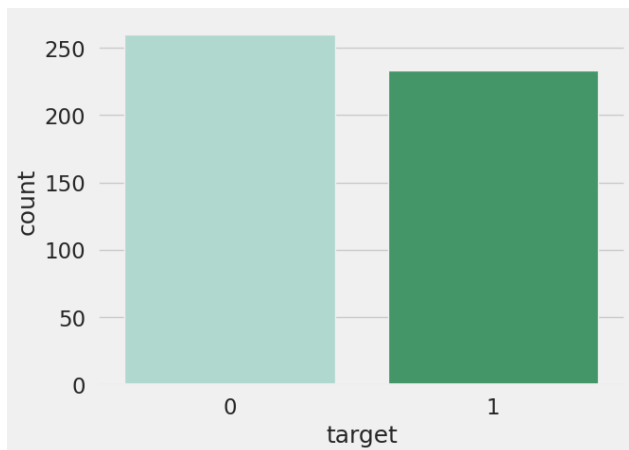


Fig. 2: Count of patients with and without heart disease

Figure 3 illustrates the distribution of the attributes, their count and the data type.

The dataset includes information from 493 patients, encompassing 13 distinct parameters. It features both male and female participants.

#	Column	Non-Null Count	Dtype
0	age	493 non-null	int64
1	sex	493 non-null	int64
2	cp	493 non-null	int64
3	trestbps	493 non-null	int64
4	chol	493 non-null	int64
5	fbs	493 non-null	int64
6	restecg	493 non-null	int64
7	thalach	493 non-null	int64
8	exang	493 non-null	int64
9	oldpeak	493 non-null	float64
10	slope	493 non-null	int64
11	ca	493 non-null	int64
12	thal	493 non-null	int64
13	target	493 non-null	int64

Fig. 3: Dataset attributes

The graph reveals that approx. 260 patients have no signs of heart disease (0), while 233 percent have (1). Although the distribution is not exactly 50:50, it is sufficiently balanced to prevent inconsistencies. Figure 4 illustrates the age distribution of the dataset, with patients free of heart disease represented in blue and those with heart disease shown in orange.

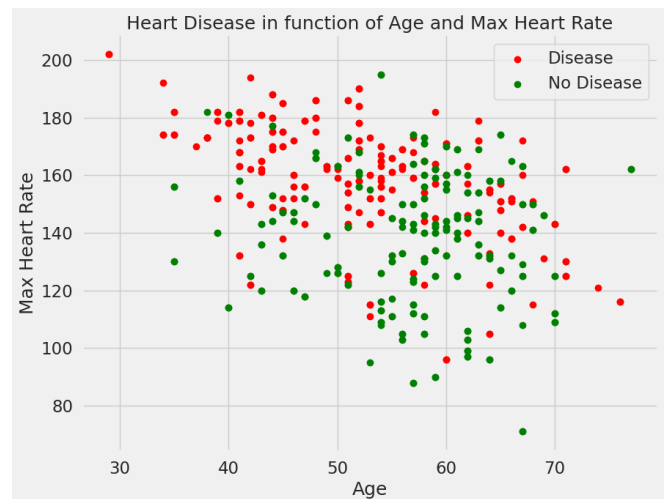


Fig. 4: Frequency of heart disease for all ages

The age group of 40 to 60 has the highest prevalence of heart diseases, whereas individuals between 60 to 70 demonstrate the lowest incidence of individuals without any heart disease.

The maximum heart rate attained by a patient who is 30 years old exceeds 200 beats per minute. In general, individuals who

have elevated heart rates have been clinically diagnosed with cardiovascular disorders.

Figure 5 illustrates the relationship between sex and the occurrence of heart disease. The graph shows that the number of heart disease cases is more for females. However, male patients are having disease for almost all age groups.



Fig. 5: Male Vs female according to age

A histogram displays data points in user-defined ranges in a graphical format. Figure 6 shows a histogram of the features in the UCI Dataset.

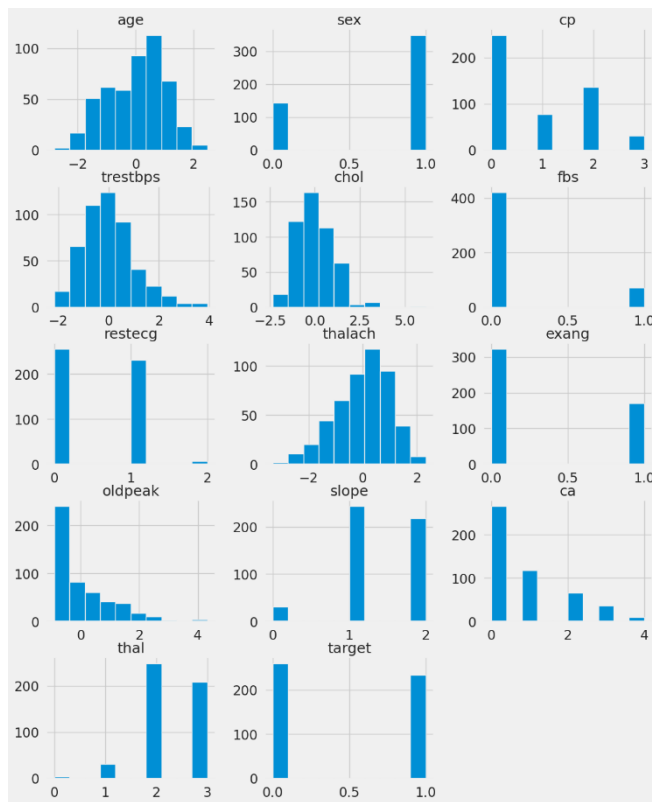


Fig. 6: Dataset histogram

The histogram is a visual representation that resembles a bar graph and serves to succinctly present a dataset by organizing several data points into coherent intervals or bins.

V. MODEL IMPLEMENTATION

This study utilizes models developed with KNN and SVM algorithms. Any missing or erroneous values in the dataset have been eliminated. The data was divided into a 70:30 ratio for training and testing purposes.

In supervised machine learning, the process involves observing how various inputs and outputs are utilized to learn a function that connects inputs to outputs. This function is derived from the labels assigned to a collection of training instances. Supervised algorithms, which are a category of machine learning algorithms, require human supervision to ensure accurate results. Input is gathered from both a "train" dataset and a "test" dataset. The goal is to predict or classify an output variable within the training dataset. Each algorithm identifies patterns from the training data before making predictions or classifications. The process of supervised learning is illustrated in Figure 7.

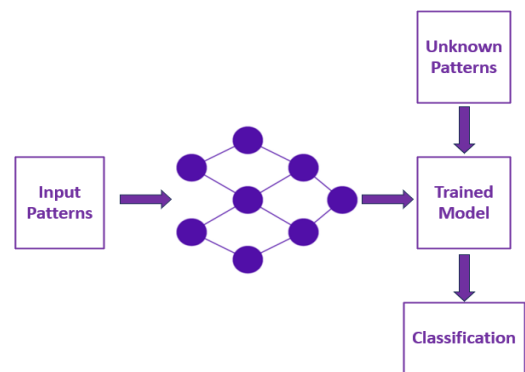


Fig. 7: Supervised learning

VI. RESULTS ANALYSIS

The performance of a classification model is evaluated using its confusion matrix. The True Negative value, located in the top left corner, indicates that both the observed and predicted values were negative. The False Positive value, found in the top right corner, signifies that the actual value was "No," but the model incorrectly predicted "Yes," which is also referred to as a Type I error. The False Negative value, often called a Type II error, is represented in the bottom left corner; if this value is True, it means the model's prediction was False. Finally, when both the actual value and the model's prediction match, as indicated by the number in the bottom right corner, we refer to this as a True Positive.

The confusion matrix for the KNN implementation is given in figure 8.

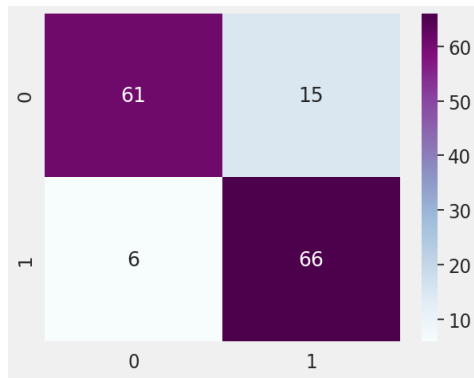


Fig. 8: KNN confusion matrix

In the figure 8, the (0,0) cell represents that in the test pattern 61 patients were having no disease and the model also recognized correctly. However, 6 patients in the cell (0,1) in the test data were having no disease but the model misclassified as these patients are having disease. Similarly, the cell (1,1) represents the patients were having heart disease and the model also forecast them as having heart disease whereas the cell value (1,0) signifies the patients were having heart disease but the model predicted as having no disease. The correct prediction is calculated as 86% and misprediction is calculated as 14%. The classification report is given in figure 9.

	precision	recall	f1-score	support
0	0.91	0.80	0.85	76
1	0.81	0.92	0.86	72
accuracy			0.86	148
macro avg	0.86	0.86	0.86	148
weighted avg	0.86	0.86	0.86	148

Fig. 9: KNN classification report

In the similar way, the confusion matrix has been generated for the implementation of SVM model.

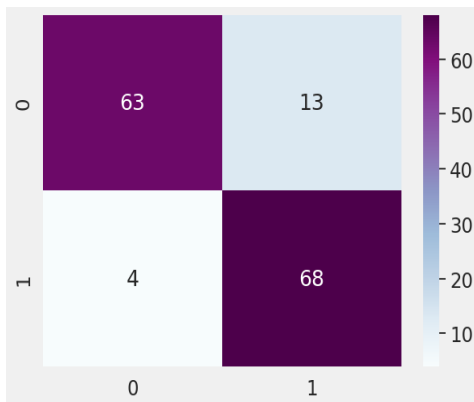


Fig. 10: SVM confusion matrix

In the figure 10, the (0,0) cell represents that in the test pattern 63 patients were having no disease and the model also recognized correctly however, 4 patients in the cell (0,1) in the test data were having no disease but the model misclassified as these patients are having disease. Similarly, the cell (1,1) represents the patients who were having heart disease and the model also predicted them as having heart

disease whereas the cell value (1,0) represents the patients who were having heart disease but the model predicted as having no disease. The correct prediction is calculated as 89% and misprediction is calculated as 11%. The classification report is given in figure 11.

	precision	recall	f1-score	support
0	0.94	0.83	0.88	76
1	0.84	0.94	0.89	72
accuracy			0.89	148
macro avg	0.89	0.89	0.89	148
weighted avg	0.89	0.89	0.88	148

Fig. 11: SVM classification report

The table below illustrates a significant variation in accuracy scores among different Machine Learning Algorithms. SVM has been selected as the optimal algorithm for the model due to its highest accuracy score, making it effective for predicting the development of cardiac issues.

Table 2 summarizes the algorithms used and the model accuracy.

TABLE-2: ACCURACY COMPARISON		
Algorithms	Accuracy	Features
SVM	89%	Linear Kernel
KNN	86%	8 Nearest Neighbors

VII. CONCLUSION

According to medical reports, heart disease now a days is a major cause of death worldwide claiming millions of lives each year. Prompt diagnosis is essential when a heart condition is suspected, as it allows for timely treatment. Additionally, medications are available, and some patients may require surgical intervention if lifestyle changes alone are insufficient.

Currently, there is no universally accepted dataset for predicting cardiac diseases. In this study, the authors have developed two effective models utilizing a diverse array of datasets. Ensemble models, which combine multiple machine learning algorithms, have demonstrated the highest accuracy in making predictions. The establishment of a standardized dataset is crucial for creating reliable and effective diagnostic models.

In this research, the SVM model achieved an accuracy rate of 89%, surpassing the KNN model's 86%. This does not diminish the effectiveness of the KNN model, as different datasets may yield varying results. Overall, it can be concluded that machine learning models are promising tools for medical diagnosis, with significant potential for broader applications in the future.

REFERENCES

- [1] Bhardwaj, S., Jain, S., Trivedi, N., Tiwari, R. (2022). Intelligent Heart Disease Prediction System Using Data Mining Modeling Techniques. 10.1007/978-981-19-0707-4_79.
- [2] Tiwari, S., Kumar, S., & Guleria, K. (2020). Outbreak trends of coronavirus disease-2019 in India: a prediction. Disaster medicine and public health preparedness, 14(5), e33-e38.

- [3] Kumar, I., Kumar, A., Kumar, V.D.A., Kannan, R., Vimal, V., Singh, K.U., Mahmud, M., Dense Tissue Pattern Characterization Using Deep Neural Network, (2022) *Cognitive Computation*, 14 (5), pp. 1728-1751. DOI: 10.1007/s12559-021-09970-2.
- [4] Negi, P., John, D., Vimal, V., Prasad, V. Machine Learning-Based Analysis of Echocardiography Images for Cardiac Disease Diagnosis, (2023) *International Journal of Intelligent Systems and Applications in Engineering*, 11 (7s), pp. 08-14.
- [5] Gudur, A., Sivaraman, H., Vimal, V. Deep Learning-Based Detection of Lung Nodules in CT scans for Cancer Screening, (2023) *International Journal of Intelligent Systems and Applications in Engineering*, 11 (7s), pp. 20-28.
- [6] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*. 2023; 16(2):88. <https://doi.org/10.3390/a16020088>
- [7] Mohammad, F.; Al-Ahmadi, S. WT-CNN: A Hybrid Machine Learning Model for Heart Disease Prediction. *Mathematics* 2023, 11, 4681. <https://doi.org/10.3390/math11224681>
- [8] Liaqat, Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in *IEEE Access*, vol. 7, pp. 54007-54014, 2019, doi: 10.1109/ACCESS.2019.2909969.
- [9] Kaur, B., Kaur, G. (2023). Heart Disease Prediction Using Modified Machine Learning Algorithm. *International Conference on Innovative Computing and Communications*. Lecture Notes in Networks and Systems, vol 473. Springer, Singapore. https://doi.org/10.1007/978-981-19-2821-5_16
- [10] Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Nayyar, A., & Kwak, K. S. (2023). An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms. *Computer Systems Science & Engineering*, 46(3).
- [11] Ahmad, G. N., Fatima, H., & Haris, M. (2023, February). Hybrid Machine Learning Algorithms for Optimal Diagnosis of Heart Disease with Feature Analysis. In 2023 International Conference on Power, Instrumentation, Energy and Control (PIECON) (pp. 1-6). IEEE.
- [12] Palaniappan, S. & Awang, R. In 2012 IEEE/ACS International Conference on Computer Systems and Applications 108–115 (IEEE, New York).
- [13] Olaniyi, E., Oyedotun, O., Khashman, A. (2015). Heart Diseases Diagnosis Using Neural Networks Arbitration. *International Journal of Intelligent Systems and Applications*. 7. 75-82. 10.5815/ijisa.2015.12.08.
- [14] Guharoy, R., Dubey, B., Mohapatra, S.K. and Sidharth, S., 2024, March. Advanced Cardiovascular Health Management Through Comprehensive Risk Assessment and Predictive Modeling of Cardiac Arrest. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 1-5). IEEE.
- [15] Mohapatra, S.K. and Jain, A., 2023, April. Predictive Analysis of Stroke Prediction by using Machine Learning Implementations. In 2023 IEEE International Conference on Contemporary Computing and Communications (InC4) (Vol. 1, pp. 1-6). IEEE.
- [16] Sahu, P., Mohapatra, S.K., Sarangi, P.K. and Mohanty, J., 2023, February. Discrimination and Investigation of Cardiac Infarction gleaned from Various Machine Learning and Optimization Methods. In 2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET) (pp. 1-5). IEEE.
- [17] University of California, Irvine, Heart Disease Data Set, Available at: <https://archive.ics.uci.edu/ml/datasets/heart+disease>, [Accessed on 19th Feb, 2022]