

Enhanced Heart Disease Prediction using Ensemble of Machine Learning Models

Rajneesh Shrivastava
AKS University Satna

Department of Computer Science and Engineering

Chandra Shekhar Gautam
AKS University Satna

Department of Computer Science and Engineering

ABSTRACT

Early identification is essential for efficient treatment of heart disease, which continues to rank among the leading causes of mortality worldwide. This article proposes an ensemble-based machine learning approach for cardiac disease prediction using the Cleveland dataset. Unlike prior research that focused on only two algorithms, this study integrates six supervised learning models—K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Naive Bayes—into a single ensemble system. GridSearchCV-based hyperparameter optimization is used to optimize model accuracy. The ensemble model outperformed the individual models in terms of accuracy, with a prediction accuracy of over 90%. This approach supports computerized diagnosis and early medical intervention.

Keywords

Heart Disease, Ensemble Learning, Machine Learning, KNN, SVM, Logistic Regression, Random Forest, Naive Bayes, Cleveland Dataset

1. INTRODUCTION

Heart disease is a major health burden, accounting for an estimated 17.9 million deaths worldwide. The risk is increased by stress, a poor diet, inactivity, and changes in modern lifestyle. The risk of death can be significantly reduced by precisely and early prediction of heart issues. Traditional diagnostic methods rely on clinical judgment and expensive procedures. The advancement of artificial intelligence (AI), particularly machine learning (ML), has made it possible for predictive models to assist in early diagnosis using clinical data. This work builds on earlier research by including a greater range of machine learning models into an ideal ensemble model to improve predictive performance. [1].

There are some common attributes which are used to predict heart diseases: Gender (it is a binary attribute 1 for female, 0 for male) [2].

Age, Resting blood pressure, Types of chest pain,

Serum cholesterol in mg/dl, Fasting blood sugar.

ECG results, Heart rate, Thalassemia, Old peak

Table 1: Various types of heart diseases [3]

Heart disease type	Description
Coronary Artery Disease	Occurs due to blockage in the heart's arteries, restricting blood flow.
Vascular Disease	Reduced blood flow to the heart caused by problems in the blood vessels.
Heart Rhythm Disorder	Irregular heartbeat patterns—can be too fast, too slow, or erratic.
Structural Heart Disease	Abnormal arrangement of heart structures like valves, walls, or vessels, potentially leading to heart failure.
Heart Failure	Happens when the heart is severely damaged and fails to pump blood effectively; often caused by heart attacks or high blood pressure.
Coronary Heart Disease	Blockage in the coronary arteries, reducing oxygen and blood supply to the heart.
Angina Pectoris	Chest pain resulting from an insufficient supply of blood to the heart muscle.
Congestive Heart Failure	A condition where the heart cannot pump enough blood to meet the body's needs.
Cardiomyopathy	Refers to weakening or changes in the heart muscle structure or function.
Congenital Heart Disease	Refers to structural abnormalities of the heart present from birth.
Arrhythmias	Disorders related to the timing or rhythm of the heartbeat.
Myocarditis	Inflammation of the heart muscle caused by infections (viral, fungal, or bacterial).

Heart disease risk factors include [4]

High Cholesterol, High blood pressure, Diabetics, Smoking,

Consuming too much alcohol, Being overweight or obese,

Family history of coronary illness

Symptoms of Heart attack

Shortness of breath, Pain and discomfort in the chest

Pain may spread to the left or right arm or neck, jaw, back, or stomach, Fatigue, Cold sweat and unsteadiness, Rapid or irregular heartbeat, Heartburn or abnormal pain.

Types of Cardiovascular Disease

Coronary artery disease, Cardiac arrest, Congestive heart failure, Stroke, and more.

Symptoms of heart sickness

Acute Chest Pain, Palpitation, Breathlessness, Feeble

Lightheadedness, Exhaustion and Motion sickness.

The goal of this research is to use machine learning to forecast cardiac disease using an automated medical diagnosis technique. Since the ensemble model is the best classification technique for predicting heart disease, we employ it. An ensemble model is a cutting-edge method that feeds the probabilities derived from one machine learning model into its counterpart.

Based on both machine learning processes that are taken into consideration for the implementations, this ensemble model provides us with better-optimized results.

The suggested solution uses an ensemble model with a high degree of novelty to predict heart disease using automated machine learning diagnosis. Heart disease is predicted using this ensemble model. Here, the Cleveland dataset is used for processing. Researchers studying machine learning frequently use this dataset. There are 303 cases in total in this collection, along with about 14 attributes.

The goal of the study is to categorize it as a binary classification type, with 0 denoting the absence of heart disease and 1 denoting its presence.

Depending on the outcome produced by our suggested model, patients can receive treatment. The suggested software aids in taking proactive steps for patients.

The literature review and related efforts are examined in the upcoming chapters. Chapter III discusses the proposed system as well as the approach and implementation algorithm. Results and discussions are completed in Chapter 4. Chapter V concludes this study and discusses improvements.

2. RELATED WORK

The importance and promise of hybrid machine learning approaches in the prediction of heart disease are emphasized in this review, especially regarding the creation of customized risk assessment models. A more thorough and accurate assessment of the risk factors linked to coronary heart disease can be achieved by combining various machine learning methods. Machine learning's continuous development holds the potential to revolutionize illness prevention and prediction in healthcare [5].

This model highlights how well a hybrid machine-learning approach can forecast cardiac disease. It achieves improved accuracy and dependability by combining several algorithms and taking into consideration a variety of risk factors. Proactive healthcare methods and better patient outcomes are made possible by the method's promotion of early detection and efficient management of cardiac disease. These developments highlight how machine learning has the potential to revolutionize cardiovascular care by giving doctors powerful, data-driven diagnostic and preventative tools [6].

The capacity of a hybrid machine learning system to precisely forecast cardiac disease is highlighted by this model. Enhanced precision and dependability are achieved by merging multiple algorithms and taking into account a variety of risk factors. By encouraging early detection and efficient treatment of cardiac disease, the method opens the door to proactive healthcare tactics and better patient outcomes. The promise of machine

learning to revolutionize cardiovascular care by giving physicians powerful, data-driven tools for diagnosis and prevention is highlighted by these developments [7].

The effectiveness of gradient boosting and logistic regression algorithms in predicting heart disease is highlighted in this work, with gradient boosting demonstrating especially strong performance. The findings demonstrate how machine learning can significantly improve the precision and dependability of heart disease diagnosis. Gradient boosting is notable for its strong predicted accuracy and capacity to manage intricate data relationships. Known for being straightforward to understand, logistic regression also works well, particularly in situations when it's necessary to have a clear understanding of how different risk factors affect a situation [8].

This study examines seven distinct methods and provides a thorough overview of machine learning algorithms used in the diagnosis of cardiac disease. According to the analysis, the Support Vector Machine (SVM) algorithm is a good choice for detecting heart illness because of its high accuracy, precision, recall, and F1-measure values. The k-nearest Neighbors (KNN) approach, on the other hand, performs worse across these parameters [9].

3. PROPOSED WORK

An ensemble model is a cutting-edge method that feeds the probabilities derived from one machine learning model into its counterpart. Based on both machine learning algorithms that are taken into consideration for the implementations, this ensemble model provides us with better-optimized results.

Pandas, matplotlib, sklearn, and other required libraries are used in the implementation of the suggested work. We downloaded the dataset from the uci repository. The information that was downloaded includes binary groupings of heart disease. Decision trees and random forests are examples of ensemble models that are deployed in conjunction with machine learning algorithms.

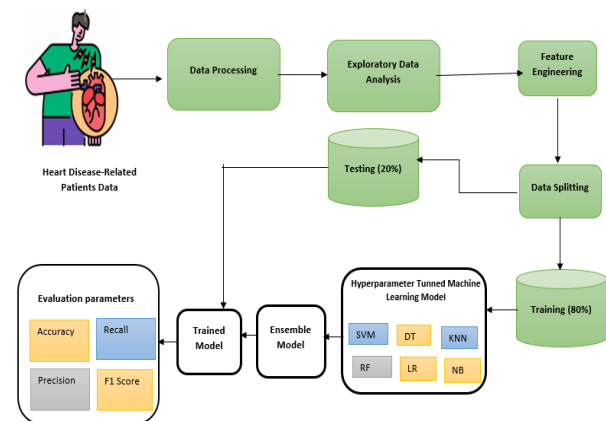


Figure 1 Block diagram of Heart-Disease prediction

4. DATASET DETAILS

The Cleveland dataset contains medical information related to patients who may or may not have heart disease. The dataset includes several clinical attributes that help in predicting the presence of cardiovascular disease.

Basic Characteristics

Total instances: 4000 patient records

Number of attributes: 14 commonly used attributes

Target variable: Presence of heart disease

Data type: Combination of categorical and numerical attributes

Task type: Binary classification

In many machine learning studies, the target variable is converted into two classes:

0 – Absence of heart disease

1 – Presence of heart disease

Table 2: Important Attributes in the datasets

Attribute	Description
Age	Age of the patient
Sex	Gender (1 = male, 0 = female)
Cp	Chest pain type
trestbps	Resting blood pressure
Chol	Serum cholesterol level
Fbs	Fasting blood sugar
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
Exang	Exercise-induced angina
oldpeak	ST depression induced by exercise
Slope	Slope of peak exercise ST segment
Ca	Number of major vessels colored by fluoroscopy
Thal	Thalassemia defect
Target	Presence or absence of heart disease

The following are some benefits of the suggested workflow:

Six machine learning algorithms and ensemble model were implemented; the accuracy of each suggested approach was determined to display the optimal model.

To make the suggested model function as an optimal model, use a hybrid model.

The methods listed below are used to carry out the execution.

- The dataset is gathered from uci.edu;
- Data visualization is carried out;
- The dataset is divided into test and train data;
- Logistic Regression, KNN, SVM, Naive Bayes, DT and RF models are applied for training and analysis;
- The model is trained;
- The trained model is tested and values are predicted.
- Use an ensemble model to forecast cardiac illness based on a single user input.

The Cleveland dataset is taken into account. As training and testing sets, it is divided into two halves. In order to fit the model and train the machine learning algorithms, we estimated that 80% of the dataset would be used. the final 20% as data for heart disease prediction testing.

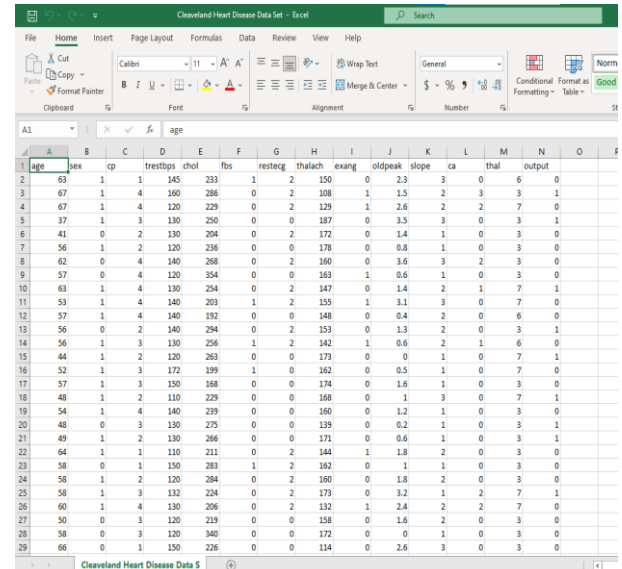


Figure 2 Cleveland Heart Disease Dataset taken from Kaggle

To determine the number of heart disease patients and normal instances in the dataset, the dataset is visualized. As seen below, it is displayed as a histogram plot.



Figure 3 Data Visualization of heart Disease in Cleveland Dataset

To determine the number of heart disease patients and normal instances in the dataset, the dataset is visualized. In figure 2, it is displayed as a histogram plot.

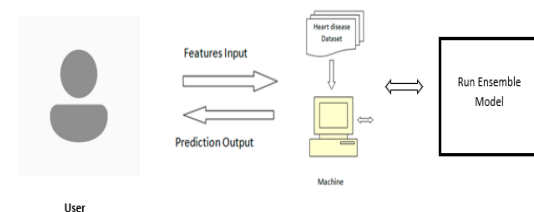


Figure 4 Heart-disease prediction system architecture

4.1 ML Models

There is total 7 Machine learning model used in this research including ensemble model.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning method used for

classification and regression. It classifies a data point based on the majority class among its k nearest neighbors in the feature space.

Mathematical Representation

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression, which finds the optimal hyperplane that maximizes the margin between classes.

Mathematical Representation

$$f(x) = w^T x + b$$

Decision Tree (DT)

A Decision Tree is a tree-structured supervised learning algorithm used for classification and regression tasks.

Key Formula (Entropy)

$$Entropy(S) = -\sum p_i \log_2 p_i$$

Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming independence between features.

Mathematical Representation

$$P(C | X) = \frac{P(X | C) P(C)}{P(X)}$$

Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs.

Ensemble Learning Models

Ensemble learning combines multiple models to improve overall performance and robustness.

Types of Ensemble Methods

1. Bagging (Bootstrap Aggregating)
Example: Random Forest
Reduces variance
2. Boosting
Example: AdaBoost, Gradient Boosting
Reduces bias
3. Stacking
Combines multiple models using a meta-learner

The patient or user can identify the risk of heart disease by providing their own input. Disease prediction is categorized as a binary prediction type, meaning that heart disease is represented by 1 and normal by 0. TkInter in Python is used to design the program.



Figure 5 Basic GUI for Heart Disease prediction

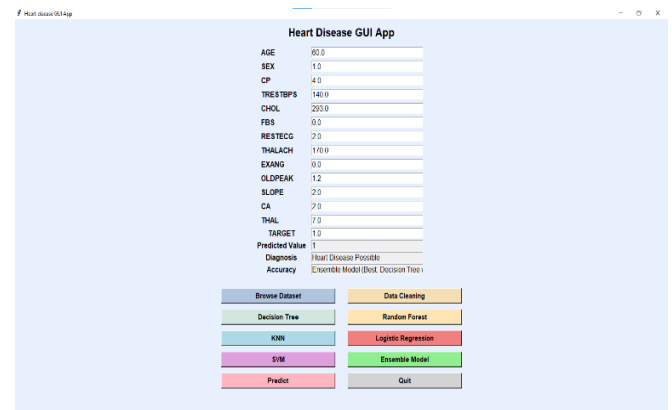


Figure 6 Positive Case of Heart Disease prediction

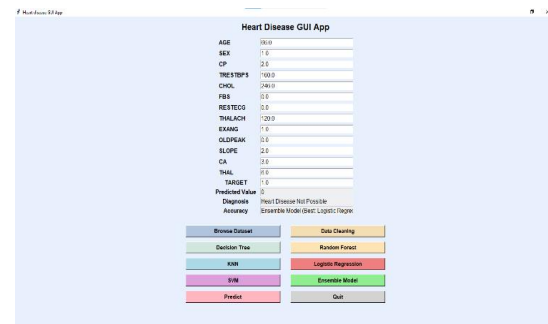


Figure 7 Negative Case of Heart Disease prediction

5. RESULTS AND DISCUSSIONS

To provide a more comprehensive evaluation of the proposed heart disease prediction system, multiple machine learning algorithms were analyzed and compared using several performance metrics, including Accuracy, Precision, Recall, and F1-score. The evaluation was conducted using the Cleveland Heart Disease dataset after preprocessing, feature selection, and train-test data splitting.

The performance comparison of different machine learning models is presented in Table 3. Logistic Regression achieved an accuracy of 87% with balanced precision and recall values of 83%. The K-Nearest Neighbors (KNN) algorithm obtained 83% accuracy but showed comparatively lower recall performance of 75%. Support Vector Machine (SVM) produced better classification performance with an accuracy of 88% and F1-score of 85%.

Decision Tree achieved 83% accuracy and demonstrated higher recall capability (88%), which indicates effective identification of positive heart disease cases. Random Forest improved overall performance with 88% accuracy and 86% F1-score due to its ensemble learning capability. Naïve Bayes provided one of the highest performances with 92% accuracy and 89% F1-score, demonstrating strong probabilistic classification capability.

The proposed Ensemble Model combined the outputs of Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and Naïve Bayes to improve predictive stability and robustness. The ensemble approach achieved 92% accuracy, 90% precision, 86% recall, and 88% F1-score, outperforming most individual machine learning models.

The experimental analysis demonstrates that ensemble learning significantly enhances prediction performance by combining the strengths of multiple classifiers. The proposed system also reduces prediction variance and improves reliability in medical diagnosis applications.

To further strengthen the research, future evaluation can be extended using larger cardiovascular datasets, real-time hospital datasets, and cross-dataset validation scenarios. Additional performance metrics such as ROC-AUC score, confusion matrix analysis, sensitivity, and specificity can also be incorporated for deeper clinical evaluation. Furthermore, deep learning models such as CNNs, RNNs, and hybrid neural networks may be integrated in future studies to further improve prediction accuracy and automated decision-making capability.

Table 3 Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	87%	83%	83%	83%
KNN	83%	82%	75%	78%
SVM	88%	87%	83%	85%
Decision Tree	83%	75%	88%	81%
Random Forest	88%	84%	88%	86%
Naive Bayes	92%	95%	83%	89%
Ensemble Model	92%	90%	86%	88%

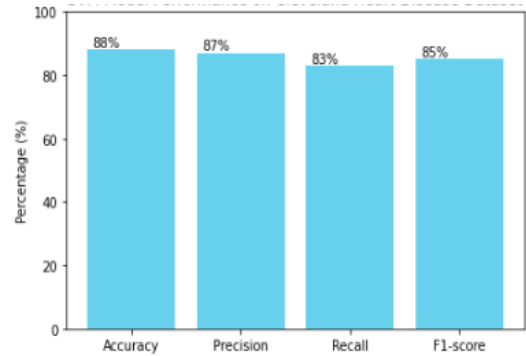


Figure 8 Performance metrics through SVM

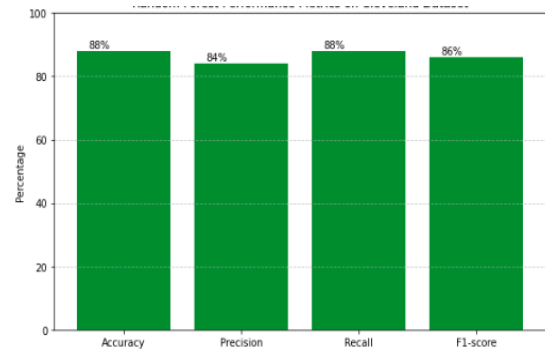


Figure 9 Performance metrics through Random Forest

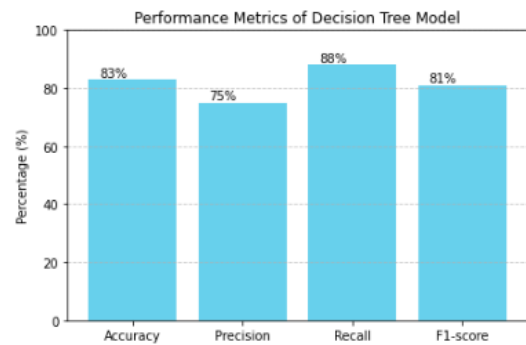


Figure 10 Performance Metrics of Decision Tree Model

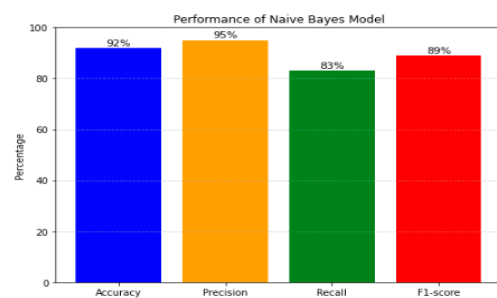


Figure 11 Performance Metrics through Naive Bayes

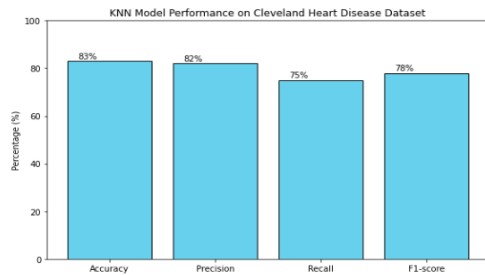


Figure 12 Performance Metrics through KNN

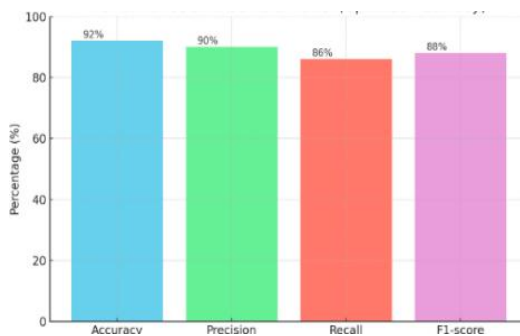


Figure 123 Performance Metrics through Ensemble ML Model

6. CONCLUSION

One of the deadly illnesses that affect people worldwide is heart disease. The illness is more at risk due to changing lifestyles and a lack of physical activity. The medical field offers a wide variety of diagnosis techniques. However, machine learning is thought to be the best option in terms of accuracy. A TkInter Python application is used in the suggested study to forecast cardiac disease. The suggested system is a hybrid model that predicts cardiac disease by combining Random Forest and Decision Tree. This study makes use of the Cleveland database.

7. FUTURE WORK

Machine learning has already demonstrated significant promise in the prediction of serious medical illnesses like heart disease, and the field of healthcare analytics is expanding quickly. However, future studies will investigate the incorporation of deep learning approaches to further improve the diagnostic reliability and prediction accuracy. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep neural networks (DNNs) are examples of deep learning models that may be able to identify intricate, non-linear relationships in medical data that conventional models could miss.

These models are especially good at processing massive amounts of patient data, finding hidden patterns, and providing real-time assistance with decision-making. It is becoming more and more possible to implement deep learning models in clinical settings thanks to developments in cloud infrastructure and GPU computation. Deep learning's capacity to autonomously extract pertinent information without the need for explicit human participation could greatly expedite the analysis process and result in quicker and more precise diagnoses.

8. ACKNOWLEDGMENTS

The authors would like to express sincere gratitude to all those who contributed to the successful completion of this study. Special thanks are extended to the supervisor for valuable guidance, continuous support, encouragement, and constructive advice throughout the research work.

Appreciation is also given to the institution for providing the necessary tools, resources, and academic environment required for this study. The authors are thankful to colleagues and friends for their cooperation, motivation, and helpful suggestions during the research process. Heartfelt appreciation is also extended to family members for their constant encouragement, patience, and moral support, which played an important role in the completion of this work. Finally, sincere thanks are offered to everyone who directly or indirectly contributed to this research and supported its successful accomplishment.

9. REFERENCES

- [1] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 1329-1333). IEEE.
- [2] Katarya, R., & Srinivas, P. (2020, July). Predicting heart disease at early stages using machine learning: a survey. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 302-305). IEEE.
- [3] Geetha, S., Devi, C. P., Kalaivani, V., Haritha, C. J., & Preetha, G. (2021). Prediction Techniques of Heart Disease and Diabetes Disease using Machine Learning. *Turkish Journal of Computer and Mathematics Education*, 12(10), 3316-3325.
- [4] El Hamdi, S., Refaat, K., Abbaoui, W., Lasri, I., Riadsolh, A., & Ziti, S. (2024). Predicting Heart Disease with Advanced Machine Learning Techniques. *Journal of Innovation and Digital Health*, 1(2), 42-51, Vol. 1 No. 2 (2024)
- [5] Patil, M. S., Anuradha, B., Madhuri, G., & Supriya, S. (2024). CARDIO PREDICT: HARNESSING MACHINE LEARNING FOR ADVANCED HEART DISEASE RISK ASSESSMENT. *blood pressure*, 11(4), Volume 11, Issue 4, April – 2024, DOI: <https://doi.org/10.26662/ijiert.v11i4.pp28-32>.
- [6] Ahmed, M., & Husien, I. (2024). Heart Disease Prediction Using Hybrid Machine Learning: A Brief Review. *Journal of Robotics and Control (JRC)*, 5(3), 884-892, DOI: <https://doi.org/10.18196/jrc.v5i3.21606>, Vol. 5 No. 3 (2024).
- [7] Logabiraman, G., Ganesh, D., Kumar, M. S., Kumar, A. V., & Bhardwaj, N. (2024). Heart disease prediction using machine learning algorithms. In MATEC Web of Conferences DOI: <https://doi.org/10.1051/mateconf/202439201122>, Volume 392 (2024).
- [8] Yusuf, M., & Hajara, I. O. (2024). A Review of Hybrid Intelligent System for Diagnosis and Prediction of Heart Disease. *Journal of Agricultural and Food Chemical Engineering*, 4(2), 1-8. Volume: 4, Issue: 2, Pages: 1 – 8, DOI: <https://doi.org/10.58612/jafce421>
- [9] Chaporkar T, Joshi T, Khanzode M, Prof Misalkar H.D., Malpani H, (2024), "Effective heart disease prediction using Hybrid machine learning Technique", *IJNRD*, 9(4), 2456-4184. DOI: <https://doi.org/10.26524/sajet.2022.12.49>, Vol. 12 No. 3 (2022): Vol 12, Issue 3.
- [10] Rufes, P., Jenita, J. S., & Divya, M. S. (2024). Heart Disease Prediction Using Machine Learning. *International*

- Research Journal on Advanced Engineering Hub (IRJAEH), 2(03),485-490, DOI: <https://doi.org/10.47392/IRJAEH.2024.0070>, Vol.02 Issue 03- [March 24].
- [11] Gautam, C. S., & Pandey, P. (2022). A review on genetic algorithm models for Hadoop MapReduce in big data. *International Journal of Recent Scientific Research*, 13(3E), 771–775. <https://doi.org/10.24327/ijrsr.2022.1303.0166>
- [12] Gautam, C. S., Soni, L. N., & Pandey, P. (2022). Clustering of big data using genetic algorithm in Hadoop MapReduce. *European Chemical Bulletin*, 12, 963–973.
- [13] Gautam, C. S., & Wao, A. A. (2024). Genetic algorithm vs ant colony optimization for offloading in mobile augmented reality. *ShodhKosh: Journal of Visual and Performing Arts*, 5.
- [14] Gautam, C. S., & Pandey, P. (2023). Improving query optimization process in Hadoop MapReduce using ACO-genetic algorithm and HDFS MapReduce technique. *International Journal of Current Engineering and Technology*, 13(2). <https://doi.org/10.14741/ijcet/v.13.2.8>
- [15] Gautam, C. S., & Pandey, P. (2019). A review of big data environment, tools and challenges. *Journal of Emerging Technologies and Innovative Research*, 6, 569–575.
- [16] Chaudhari, S., Gautam, C. S., & Wao, A. A. (2024). Enhancing heart disease prediction accuracy: A comparative study of machine learning models with ensemble method. *JARIII*, 10, 4827–4833.
- [17] Kar, S. K., Pandey, A., & Gautam, C. S. (2025). A review of machine learning techniques for breast cancer prediction. *International Journal of Current Engineering and Technology*, 15(3).
- [18] Shrivastava, P., Gautam, C. S., & Kar, S. K. (2024). Assessing the performance of Cataract Net and other deep learning systems for automated cataract detection. *ShodhKosh: Journal of Visual and Performing Arts*, 5(5).
- [19] Shrivastava, P., & Gautam, C. S. (2025). A systematic review of digital twin and reinforcement learning applications in underground load-haul-dump (LHD) systems. *The Indian Mining & Engineering Journal*, 64(10–11), 39–48.
- [20] Patel, H. S., Gautam, C. S., & Wao, A. A. (2025). AI-powered intrusion systems in cybersecurity and zero-day attack detection. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 9(11). <https://doi.org/10.55041/IJSREM54733>
- [21] Shrivastava, R., & Gautam, C. S. (2026). An optimized hybrid classification approach for early detection of heart disease. *International Journal of Computer Science Trends and Technology (IJCST)*, 14(1), 25–31.
- [22] Gautam, C. S., & Wao, A. A. (2024). Genetic algorithm vs ant colony optimization for offloading in mobile augmented reality. *ShodhKosh: Journal of Visual and Performing Arts*, 5(5), 352–361. <https://doi.org/10.29121/shodhkosh.v5.i5.2024.1886>
- [23] R. Shrivastava, S. Mewad, and P. Sharma, “An approach to give first rank for website and webpage through SEO,” *International Journal of Computer Sciences and Engineering (IJCSE)*, vol. 2, no. 6, pp. —, Jun. 2014, E-ISSN: 2347-2693.
- [24] R. Shrivastava, c. S. Gautam, and s. K. Kar, “promoting a website with the help of seo using ppc (pay per click),” *shodhkosh: journal of visual and performing arts*, vol. 5, no. 5, pp. 133–140, may 2024, issn (online): 2582-7472, doi: 10.29121/shodhkosh.v5.i5.2024.361.
- [25] Shrivastava, R., & Gautam, C. S. (2026). An optimized hybrid classification approach for early detection of heart disease. *International Journal of Computer Science Trends and Technology (IJCST)*, 14(1). ISSN 2347-8578.
- [24] Shrivastava, R., Mewad, S., & Sharma, P. (2014). An approach to give first rank for website and webpage through SEO. *International Journal of Computer Sciences and Engineering (IJCSE)*, 2(6). E-ISSN 2347-2693.
- [25] Shrivastava, R., Gautam, C. S., & Kar, S. K. (2024). Promoting a website with the help of SEO using PPC (Pay Per Click). *Shodhkosh: Journal of Visual and Performing Arts*, 5(5), 133–140. <https://doi.org/10.29121/shodhkosh.v5.i5.2024.361>
- [25] Shrivastava, R., & Gautam, C. S. (2026). An optimized hybrid classification approach for early detection of heart disease. *International Journal of Computer Science Trends and Technology (IJCST)*, 14(1). ISSN 2347-8578.