RESEARCH PAPER

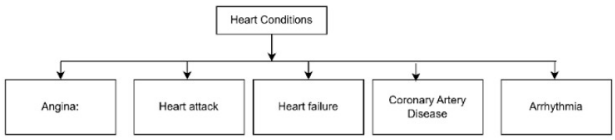
Machine Learning-Based Heart Disease Prediction: Mitigation of Overfitting on a Large Data Set

[1]Asha Hemaush Nagoriya, [2]DR. Hemangini Patel
[1]Vanita Vishram women's university, Surat, Gujarat
Pune,India. Email id : ahnagoriya@gmail.com
[2]*Computer Science,* Vanita Vishram women's university, Surat,Gujarat
Email id : hemangini.patel@vvwusurat.ac.in

Abstract— One of the biggest problems the world is currently experiencing is heart disease. In the area of clinical knowledge analysis, disorder prediction could be a significant problem. The enormous volumes of data generated by the healthcare sector have demonstrated the value of machine learning (ML) in predicting and decision-making. In this study, we present a unique approach to improve the accuracy of cardiovascular disease prediction by identifying important variables using machine learning techniques. This paper uses Python programming and machine learning approaches to forecast heart illness. We used HRFGLM algorithms on the 70000 Kaggle dataset. We use HRFGLM techniques to overcome the overfitting issues. Noise, a limited training set, and classifier complexity all contribute to overfitting. This study will look at overfitting from the standpoints of causes and remedies. We use RFE techniques to decrease dimensionality. With HRFGLM, we achieve an accuracy of 75.19% after addressing overfitting issues.

Keywords: *overfitting issues, Prognosis of heart illness, prediction of cardiovascular disease, feature selection, machine learning techniques*

How to cite this article: 1Asha Hemaush Nagoriya 2DR. Hemangini Patel. Machine Learning-Based Heart Disease Prediction: Mitigation of Overfitting on a Large Data Set. Int J Drug Deliv Technol. 2026;16(90s):105-109. DOI: 10.25258/ijddt.16.90s.11.



I. INTRODUCTION

Heart disease is currently a major worldwide health issue. [1] A variety of conditions affecting the heart and blood arteries are referred to as cardiovascular diseases (CVDs). Numerous methods effectively address the root causes of overfitting to lessen it.By stopping trainingrs. before performance on validation data declines, early stopping avoids excessive optimisation on training noise. Network reduction removes noisy data points from the training set, whereas data expansion increases datasets to better fit complex models and modify hyperparamete.Regularisation uses techniques like feature selection to penalise excessively complex models and highlight key features for great real-world performance.[2]. As seen in fig. 1 [1], cardiac ailments include a number of significant disorders that impact the heart's function.

Fig. 1: Typical Heart Conditions

Disorders that impact heart function are included under heart issues. Heart failure (poor pumping), coronary artery disease (plaque accumulation), angina (reduced blood flow), heart attack (blocked arteries), and arrhythmia (irregular heartbeat) are the main forms. It's critical to identify symptoms like dyspnea and chest pain as soon as possible.

Early detection of symptoms such as chest discomfort and dyspnea is crucial. Overfitting, which prevents us from completely applying the models to match both observed data on the training set and unseen data on the testing set, is a major issue in supervised machine learning. Overfitting in machine learning is when a model analyzes training data too deeply and finds specific patterns and noise instead of general trends. When models perform well on training data but badly on new data, this is known as overfitting.

II METHODOLOGY

The 70,000 rows and 12 columns of the Kaggle cardiovascular dataset on patient characteristics—such as age, gender, height, weight, glucose, systolic and diastolic blood pressure, exercise, cholesterol, and alcohol—are used in Table 1. The general methodological flow is depicted in Fig. 2. We gathered the data from the Kaggle website. Subsequently, remove the id, null value, missing value, and outlier, do feature engineering, and add four new columns: BMI, MAP, pulse pressure, and bp category. The data was then split into 80% training data and 20% testing data. Overfitting Reduction Techniques in relation to training data "And mention:

1. Regularisation (L2)
2. Cross-validation
3. The choice of features
4. Adjusting hyperparameters

5. Ensemble Random Forest

On the other hand, overfit models yield accurate results for the training set but not for the test set due to their large variance. Overfitting occurs when a model fits too closely to the training dataset and is unable to generalise.

A part of the training data is typically utilised as test data to look for overfitting. A high error rate in the testing data is a sign of overfitting. One method for detecting overfitting is below-fold cross-validation. Overfitting occurs when a model fits too closely to the training dataset and is unable to generalise.

There are several reasons why overfitting happens, such as:

• The training set does not contain enough data samples to fully represent all possible input data values. The training set contains a significant amount of "noisy data," or irrelevant information.

• The model is trained on a single sample of data for an excessive amount of time. The model detects the noise in the training set because to its great complexity.

Overfitting is an undesirable machine learning behaviour that occurs when a model makes accurate predictions for training data but not for new data. Before employing machine learning models to make predictions, data scientists first train them on a known data set. The model then attempts to forecast results for fresh data sets using this information. An overfit model cannot function properly for all kinds of incoming data and may produce erroneous predictions.

Attribute	Description
Age	Objective Feature | age | int (days)
Height	Objective Feature | height | int (cm)
Weight	Objective Feature | weight | float (kg)
Gender	Objective Feature | gender | categorical code
Systolic blood pressure	Examination Feature | ap_hi | int
Diastolic blood pressure	Examination Feature | ap_lo | int
Cholesterol	Examination Feature | cholesterol 1: normal 2: above normal 3: well above normal
Glucose	Examination Feature | gluc 1: normal 2: above normal 3: well above normal
Smoking	Subjective Feature | smoke | binary
Alcohol intake	Subjective Feature | alco | binary
Physical activity	Subjective Feature | active | binary
Presence or absence of cardiovascular disease	Target Variable | cardio | binary

TABLE 1. Dataset

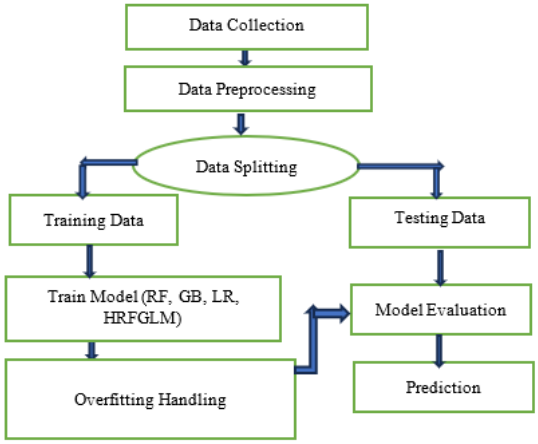


Fig. 2. Flowchart of Methodology

III LITERATURE REVIEW

A proposed stacking technique combines the outcomes of KNN, Random Forest, and SVM and employs Logistic Regression as the meta-classifier. This method yielded an accuracy of 75.1% using the Kaggle cardiovascular disease dataset, which has 70,000 patients and 12 attributes [6]. KNN, Random Forest, Naive Bayes, and Logistic Regression were among the machine learning methods evaluated. On the Kaggle cardiovascular disease dataset, which included 70,000 patients and 12 attributes, their best model produced an accuracy of 70% [7]. It achieved a 72.77% accuracy rate in predicting heart disease using a Decision Tree classifier. The Kaggle cardiovascular sickness dataset, which contains 70,000 patients and 12 attributes, was also used in the study [8]. A Random Forest classifier and a repeated random validation procedure yielded the greatest accuracy of 89.01%. The UCI cardiovascular dataset, which comprises 303 patients and 14 characteristics, was used in their tests [9]. Using a neural network for holdout cross-validation, they obtained an accuracy of 71.82% on the Kaggle cardiovascular sickness dataset, which included 70,000 patients and 12 attributes [10].It suggested using the UCI cardiac dataset to predict heart disease using a Hybrid Random Forest Linear Model (HRFLM). Their method produced an accuracy of 88.7%, indicating that prediction effectiveness was increased when Random Forest and a Linear Model were combined [11]. The Cleveland dataset was utilised to predict cardiovascular illness using HRFLM (Random Forest + Logistic Regression). Their approach outperformed traditional machine learning models in terms of interpretability, achieving a high accuracy of 98.54% [12].It developed a hybrid machine learning model that uses Random Forest, Logistic Regression, and Naïve Bayes to predict cardiac disease using the UCI Cleveland dataset. They noted improved robustness as a result of the hybrid strategy and reported 88.7% accuracy with an F1-score of 0.87 [13].The UCI dataset demonstrated enhanced prediction with effective training time when HRFLM was applied, with an accuracy of 87.7% [14]. It achieved 88% accuracy using HRFLM (Random Forest + Linear Regression) on the Cleveland dataset. They discovered that the hybrid model improved precision and generalisation [15].

IV CLASSIFICATION

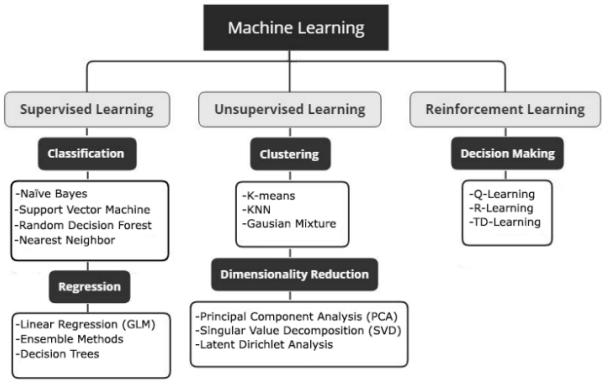


Fig. 3 Flowchart for Machine Learning Algorithms [16]

The flowchart for the machine learning algorithm is shown in Fig. 3. The three types of learning are reinforcement learning, supervised learning, and unsupervised learning.

5.1 RANDOM FOREST:

One part of supervised learning is Random Forest. Regression and classification issues can be solved with the simple and adaptable Random Forest supervised machine learning technique. It will use ensemble learning methods to construct trees. Figure 4 shows how the RF technique uses the data samples to build multiple decision trees, gathers predictions from each tree, and then uses majority voting to choose the optimal option.

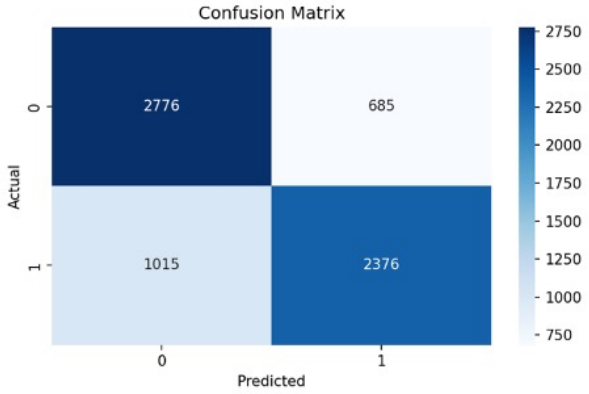


Fig. 4. Random forest [17]

5.2 LOGISTIC REGRESSION

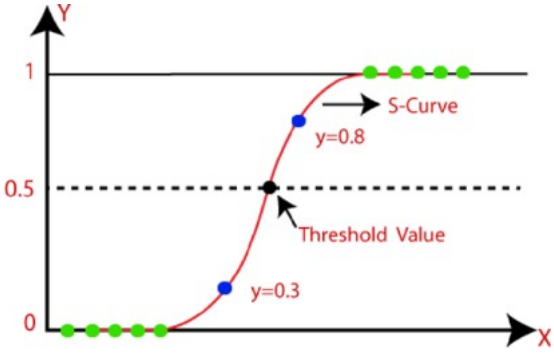


Figure 5 shows logistic regression. The sigmoid function, which accepts input as independent variables and generates a probability value between 0 and 1, is used in logistic regression for binary classification.

Fig. 5. Logistic Regression [18]

5.3 GRADIENT BOOSTING

Gradient Boosting (GB) is a potent ensemble learning method in machine learning for both regression and classification applications. It produces a single "strong" prediction model by progressively combining several "weak" models, usually shallow decision trees. It builds models one after the other with an emphasis on fixing mistakes from earlier models in order to improve performance. There are a wide variety of possible outcomes.

HRFGLM

Model of Hybrid Ensemble (HRFGLM). The Model of Hybrid Ensemble (HRFGLM) demonstrates how a soft voting classifier combines the benefits of Random Forest (RF), Gradient boosting and Logistic Regression (LR) by utilising the probability of each model. The Random Forest (RF) and Gradient boosting model's prediction are given twice as much weight as the Logistic Regression (LR) forecast because of the 2:2:1 weight ratio**.**

V.EXPERIMENTAL RESULT

The confusion matrix is displayed in Fig. 6. Your model's accuracy in classifying the two classes—0 = No Heart Problem, 1 = Heart Problem—is displayed in a confusion matrix.Results that are true positives (TP) These are actual individuals without heart disease who were correctly predicted to have a score of 0, or 2776. False Positive (FP) indicates that the anticipated 1 value is 685, but the actual class is 0.False Negative (FN) indicates that the expected 0 value is 1015, but the actual class is 1. True Positive (TP) indicates that the anticipated 1 value is 2376 and the actual class is 1..

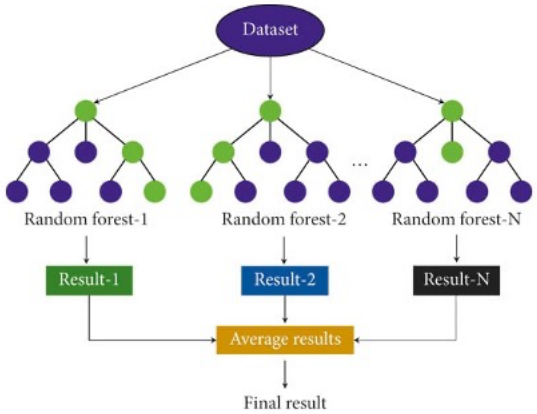


Fig. 6 Confusion matrix

Precision, recall, accuracy, and F1-score were included in the classification reports in Table 2.

Class	Precision	Recall	F1-Score	Support
No Cardio	0.73	0.80	0.77	3461
Cardio	0.78	0.70	0.74	3391
Accuracy			**0.75**	6852
Macro Avg	0.75	0.75	0.75	6852
Weighted Avg	0.75	0.75	0.75	6852

TABLE 2. classification Report

Figure 7 shows the Area Under the Receiver Operating Characteristic Curve, or AUC-ROC curve. The model distinguishes between favourable and unfavourable conditions using RFE-selected characteristics.It evaluates a classification model's ability to differentiate across classes. Reliable and steady predictions are indicated by the smooth, constantly growing form of the ROC curve.

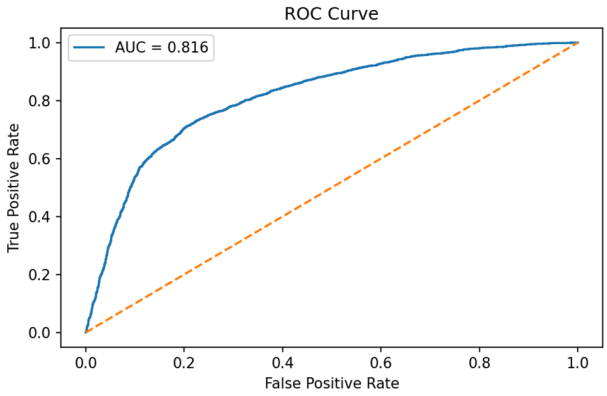


Fig. 7 AUC-ROC curv

Given that the training and testing accuracy gap is −1.01%, the overfitting analysis Fig. 8 HRFGLM model has good

generalization. There is no significant overfitting because the gap is negative and very small.

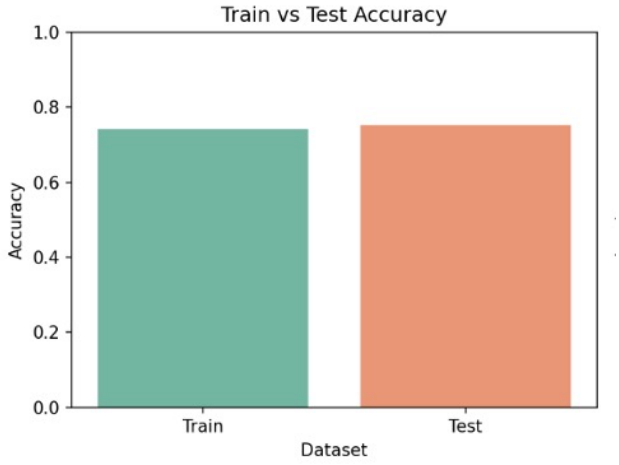


Fig.8 overfitting Analysis

Table 3 demonstrates that practical ML assessment criteria enable overfitting range thresholds.

Range		value
0.00–0.03		(No Overfitting)
0.03–0.07		(Mild Overfitting)
>0.07		(Strong Overfitting)

Table 3 Overfitting range Thresholds are supported by practical ML evaluation standards

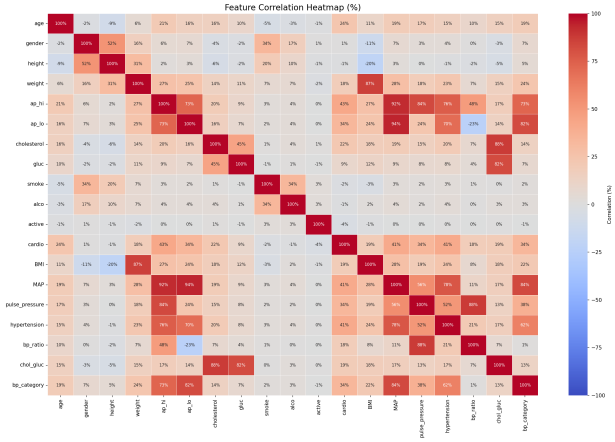


Fig. 10 Feature Correlation Heatmap

Strong positive correlations between blood pressure variables (ap_hi, ap_lo, MAP, pulse_pressure, and bp_category) are emphasised in Fig. 10 heatmap, which illustrates the relationship between characteristics associated to heart disease.

VI. CONCLUSION

Regularisation, cross-validation, and hyperparameter tuning are used in "Overfitting Handling (Reg + CV + Tuning)" to improve model performance, reduce overfitting, and boost model generalisation. The HRFGLM model improves generalisation by reducing overfitting. Some techniques that may result in overfitting include outlier removal (IQR), feature selection (RFE), SMOTE balancing + LR, hybrid voting, and soft probability weighting.

REFERENCES

[1] Kotia, A. S. S., Rastogi, M., & Bhongade, R. A. (2023). Use of machine learning techniques for effective prediction of heart disease. *CARDIOMETRY*, 315-321.

[2]Ying, X. (2019, February). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, No. 2, p. 022022). IOP Publishing.

[3] Hu, Y. (2024). Analysis and research on overfitting and underfitting issues in heart disease prediction models. *Highlights in Science, Engineering and Technology*, *123*, 116-123.

[4] Eloutouate, L., Tani, H. G., Elaachak, L., Elouaai, F., & Bouhorma, M. (2025, July). Optimizing Machine Learning for Healthcare Applications: A Case Study on Cardiovascular Disease Prediction Through Feature Selection, Regularization, and Overfitting Reduction. In *Computer Sciences & Mathematics Forum* (Vol. 10, No. 1, p. 13). MDPI.

[5] Rehman, M. U., Naseem, S., Butt, A. U. R., Mahmood, T., Khan, A. R., Khan, I., ... & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, *15*(1), 13361.

[6] Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology*, *49*(5), 600-634.

[7] Ahmed, A. M., Bataineh, B., Shakah, G., Al Enany, M. O., Aboghazalah, M. M., & Khattab, M. M. (2025). A hybrid extreme machine learning model for predicting heart disease. *Bulletin of Electrical Engineering and Informatics*, *14*(5), 4125-4137.

[8] Massari, H. E., Gherabi, N., Mhammedi, S., Ghandi, H., Bahaj, M., & Naqvi, M. R. (2024). The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. *arXiv preprint arXiv:2405.20414*.

[9].Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *J. Southwest Jiaotong Univ.* **2021**, *56*, 220–240.

[10].Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. *Int. J. Comput. Appl.* **2020**, *176*, 46–54.

[11]Sivaji, U., Rao, N. K., Srivani, C., Sree, T., & Singh, M. (2021). A Hybrid Random Forest Linear Model approach to predict the heart disease. *Annals of the Romanian Society for Cell Biology*, *25*(6), 7810-7814.

[12] Sathwik, A. S., Naseeba, B., & Challa, N. P. (2023, July). Cardiovascular Disease Prediction Using Hybrid-Random-Forest-Linear-Model (HRFLM). In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 192-197). IEEE

[13] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[14] Sweatha, M., Ramya, S., Icewariiya, S., & Kalaivani, V. (2022). Heart disease prediction using HRFLM. *International Journal of Advances in Engineering and Management (IJAEM), 4*(6), 1619-1623. https://doi.org/10.35629/5252-040616191623

[15] Khan, Z., Sikandar, G., & Anwar, S. (2023). Heart disease prediction using hybrid random forest and linear model. *International Journal of Emerging Engineering and Technology (IJEET, 2*(1), 6–12.

[16] Fri, C., & Elouahbi, R. (2021, June). Machine learning and deep learning applications in e-learning systems: A literature survey using topic modeling approach. In *2020*

6th IEEE Congress on Information Science and Technology (CiSt) (pp. 267-273). IEEE.
[17]Fu, H., & Qi, K. (2022). Evaluation model of teachers' teaching ability based on improved random forest with grey relation projection. *Scientific Programming*, *2022*(1), 5793459.
[18]Majumder, A. B., Gupta, S., Singh, D., & Majumder, S. (2021, February). An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression. In *Journal of Physics: Conference Series* (Vol. 1797, No. 1, p. 012011). IOP Publishing.