Enhancing the Prognosis of Heart Disease via Hyperparameter Optimisation of Machine Learning Models and Outlier Detection

Uzama Sadar, Parul Agarwal*, Suraiya Parveen

Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India

Life.uzma@gmail.com, pagarwal@jamiahamdard.ac.in*, suraiya@jamiahamdard.ac.in

Abstract

Information about medical research has verified that the most common cause of death and health loss nationwide is heart disease, which affects the heart and blood vessels. The number of premature deaths could be reduced by early detection of certain illnesses. Machine learning and data mining continue to be leveraged in identifying diseases based on an individual's distinct traits. The complexity of understanding the datasets' goals, the presence of too many variables to examine, and the lack of performance accuracy have, nevertheless, frequently posed difficulties for these approaches. This study presents a machine learning (ML) based approach for cardiovascular disease (CVD) prediction using integrated datasets (Cleveland and Statlog) to enhance model robustness. The proposed work addresses key challenges such as class imbalance and outlier influence, which often degrade predictive accuracy in clinical data analysis. Data preprocessing techniques, including balancing methods and outlier removal, were applied to ensure a reliable dataset. Three supervised ML classifiers, Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM), were trained and evaluated using five-fold cross-validation and RandomizedSearchCV for hyperparameter optimisation. Performance was measured employing F1-score metrics, recall, accuracy, and precision. Among the models, RF demonstrated the highest predictive accuracy of 98.2%, followed by GB at 97.6%, indicating the effectiveness of ensemble methods in heart disease prediction. The findings highlight machine learning's potential for clinical decision assistance and risk assessment in cardiovascular health.

Keywords: *Heart Disease, Machine Learning, Hyperparameter Optimisation, UCI Heart Disease dataset, Prediction.*

1. Introduction

Heart ailments rank among the most challenging conditions to treat because they occur when the heart's arteries cannot adequately pump oxygen-rich blood to other areas of the body. Heart disease affects millions of people every year and places an enormous strain on healthcare systems. 17.5 million deaths nationwide are tied to heart disease and stroke each year, according to research by the World Health Organisation[1]The diagnosis of heart disease is often made following a physical examination and observation of symptoms. Cardiovascular disease risk elements comprise smoking, ageing, genetic history, blood pressure rise, obesity, diabetes, stress, high cholesterol, and inactivity.[2]Thus, early management and precise assessment of heart disease risk have become especially crucial[3].

These days, the healthcare sector depends heavily on Artificial Intelligence (AI), Deep Learning (DL), and Machine Learning (ML) [4]. Large volumes of patient data have become available in recent years, enabling automated diagnosis of heart disease by predicting each patient's risk of developing the chronic illness using statistical methods. ML has a notable influence as a perceptive method for improving prognostic outcomes across several fields, particularly those related to heart disease risk [5][6]. AI and domain expertise are utilised to create a Decision Support System based on healthcare data[7]. Due to the intricate nature of the procedure, incorrect diagnosis or postponements in medical care could lead to increased rates of mortality and health decline. Therefore, the development of an efficient, automated, and intelligent model for predicting heart-related diseases is greatly needed [8]. Numerous research

© The Author(s), under exclusive license to Digital Manuscriptpedia. 2026 Ashok Kumar et al. (eds.), Proceedings of IMPACT-2026, DMPedia Lecture Notes in Computer Science & Engineering. ISBN: 978-81-993813-9-1

studies forecasted CVD leveraging medical datasets and Machine Learning (ML) methods. Nevertheless, class imbalance, outliers and high dimensionality in clinical datasets pose significant challenges. Therefore, not addressing these issues while using machine learning lowers the approaches' accuracy and efficiency.

The primary accomplishment of the paper is laid out as follows:

- A precise approach to forecasting coronary disease has been developed by integrating Isolation Forest, SMOTE, and machine learning algorithms.
- Isolation Forest removes outliers, and the SMOTE oversampling method balances the data.
- Experimentation is done on the combined dataset, Cleveland and Statlog.
- To validate the accuracy of the suggested model for diagnosing cardiac disease by comparing it with conventional models using several performance metrics

The document's remaining sections are organised in the following order: Section 2 highlights the literature study on the prediction of heart disease. A thorough explanation of the suggested approach, dataset description, and classification approaches, as well as materials and methods, is included in Section 3. The experimental results and discussion are covered in Section 4. The conclusion and future scope are covered in Section 5.

2. Literature Review

Heart disease, another name for cardiovascular disease, is one of the deadliest conditions that significantly raises death rates globally. However, for all doctors and researchers in the cardiovascular field, predicting such diseases has emerged as the most pressing concern. Owing to the capabilities of AI approaches, various studies have recently sought to enhance methods that use machine learning to predict heart failure and cardiovascular disorders.

Authors in the paper[9] trained different machine learning classifiers on the best characteristics of the Cleveland dataset, and produced an intelligent predictive model for the early identification of heart disease. Four feature selection techniques were used to select relevant features. Using the Extra Tree classification, 94.41% accuracy was attained. To deal with the outliers, the authors in the paper[10] have used density-based spatial clustering of applications with noise (DBSCAN) and developed a prediction model based on XGBoost.[11] First, standardized the properties of the heart-based dataset for better results. They then used the GridSearchCV approach to hyperparameter-tune machine learning classifiers. Following an evaluation of the classifier models' performance, they came to the conclusion that standardization and hyperparameter tweaking increased prediction accuracy. A 96.72% accuracy rate was attained with an SVM classifier. Using the Random Forest technique and exploratory data analysis of dataset attributes, in the paper[12] created an application for heart disease identification. An 83% accuracy rate in the training set was achieved by evaluating the significance of the cardiac dataset features using a correlation matrix. In the paper[13] The authors use AdaSyn, Particle Swarm Optimisation (PSO), and Machine Learning to develop a hybrid model for disease prediction. The recommended model outperformed with a 91.16% F1 score, 90.9% specificity, 92.8% sensitivity, 90% precision, 91.8% accuracy, and 92% area under the receiver operating characteristic curve (AUC).

In the paper[14],a new ensemble learning-based voting method was proposed using six machine learning techniques, and it was discovered that the ensemble technique outperformed all individual techniques and achieved 83%average accuracy. Similarly, in the paper [15] Researchers developed a model employing correlation feature selection and particle PSO, achieving 85.71% accuracy. Class imbalance is a problem that degrades the performance, so to deal with it, in the paper[16] The authors used the three balancing techniques and trained, tested, and assessed several classifiers, primarily focusing on optimising sensitivity for CVD risk prediction, yielding recall rates of 88% with CatBoost and SMOTE-ENN.

Since most published work focuses on the UCI online repository dataset, which is comparatively small, the authors in [17] create a hybrid dataset by merging datasets from Switzerland, Hungary, Long Beach,

and Cleveland. The CatBoost ML classifier achieves the greatest accuracy of 94.34%. Similarly, in the paper [18], the authors combined two private heart disease datasets and applied various feature selection techniques to ten ML models. XGBoost classifiers achieved 97.5% accuracy.

3. Materials and Methodology

This section presents a methodical approach pursued by dataset description, data preparation, hyperparameter optimisation, ML classification, and performance analysis.

3.1. Proposed Methodology

The authors of this study propose an integrated approach for predicting heart disease. Data collection is the initial step in which the Cleveland and Statlog heart datasets are combined to obtain more observations for the experimental study. Next, the pre-processing stage involves cleaning the data, checking for missing values, removing outliers, normalising using a Min-max scaler, and balancing using SMOTE. We then trained three distinct machine learning algorithms: Random Forest, Adaboost, and Gradient Boosting and performed a hyperparameter tuning approach to maximise our model. Lastly, by computing the various performance metrics, the efficacy of models has been evaluated. Figure1. shows our methodology's complete workflow diagram.

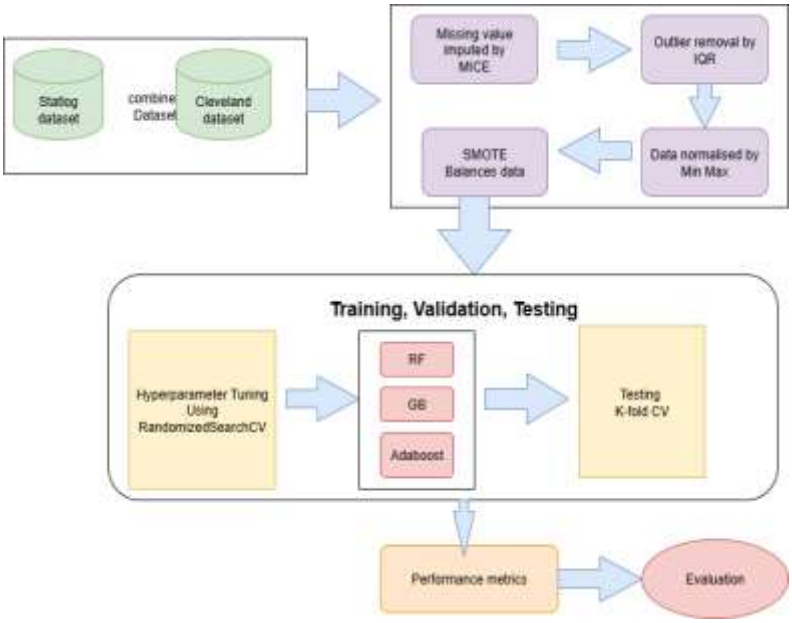

Figure 1: Complete methodology workflow

3.2. Dataset Collection

The suggested method creates 573 observations by combining two datasets from the UCI machine-learning library: Cleveland[19], having 303 observations, and Statlog[20] 270 observations total, of which 314 have no disease and 259 have disease. Table1. Shows features of the combined dataset.

Table1. Combined heart disease dataset features

Index No	Features	Characteristics
1	Age	Patients age {29-77}

2	sex	Gender of Patient{1- Male, 2-Female}
3	Cp	chest pain characteristics (typical angina-1, 2- atypical angina, 3- non-angina pain, 4-asymptomatic
4	trestbps	Blood pressure resting{94 to 200} measured in mm/ Hg
5	chol	Cholesterol serum{126 to 564} measured in mg dl
6	fbs	Blood sugar during fast> 120 mg/dl
7	restecg	Electrograph at resting 0:Normal, 1:Abnormal, 2:Hypertrophy
8	thalach	Max heart rate{71 to 202}
9	exang	Induced angina by exercise{1:yes,0:no}
10	oldpeakslope	Exercise-induced ST depression in comparison to rest{Upslope: 1, Flat: 2, downslope:3}
11	slope	The peak workout ST segment's slope{0-6.2}
12	ca	Number of fluoroscopy-colored main vessels (0–3)
13	Thal	The condition of the heart {3: normal, 6: fixed defect, 7: reversible defect}.
14	target	1:Disease, 0:no disease

3.3. Data preprocessing

Data preprocessing is the process of converting data into a precise, comprehensible format to improve the accuracy and effectiveness of models. Medical data often contains unnecessary information, is imprecise, lacks attribute values, and is incomplete [21].The Cleveland dataset, however, had six missing data two for the heart rate (Thal) attribute and four for the number of major vessels (Ca) attribute, even though both characteristics had fewer than 5% missing values, whereas the Statlog dataset doesn't have missing values.

3.3.1 Missing value imputation by Multivariate imputation by chained equations algorithm (MICE). This method was used to impute missing values in the Cleveland dataset. Imputation is carried out repeatedly by this algorithm. It is assumed that data is absent at random. This approach makes use of a regression model to forecast the value of the missing feature based on the dataset's remaining attributes[22].

3.3.2 Outlier removal by Interquartile Range (IQR). A number that deviates by a factor of three or more from the mean is called an outlier[23]. Before employing a model for prediction, identifying, removing, and eliminating outliers can significantly reduce errors and increase accuracy. Finding a

continuously distributed outlier in a dataset is made easier by the IQR approach. The data points are said to be farther apart the higher the IQR, and closer to the mean the lower the IQR. The first or lower quantile Q1 is subtracted from the third or upper quantile Q3 to find QR. In addition to taking the 50% middle point, IQR = Q3-Q1[24].After outlier removal, 408 samples are left in the dataset.

3.3.3 Data normalisation. After removing the outliers, the min-max was used to normalise the data.

3.3.4(SMOTE) Synthetic Minority Oversampling Technique It is an oversampling method that addresses an imbalanced class in a dataset. Using the Euclidean distance from its closest neighbor, SMOTE generates duplicate value of the minority class at random[25] Following the removal of outliers, 237 values were left in the majority class (disease, target=1) and 171 values in the minority class (no disease, target=0), total of 408 samples. In this case, duplicate samples were created for the minority class to bring its count to 237, matching the majority class's count. This resulted in a balanced dataset with 237 data for both classes, a total of 474 samples

3.4. ML Classifiers

3.4.1 Random Forest(RF) is a set of tree predictors that, during training, creates multiple decision trees and selects the class by voting on each tree separately. The technique randomly selects attribute locations and creates a decision tree; however, it is conceptually related to decision tree algorithms[26]. Its primary benefit is that it can increase forecast accuracy without raising processing costs.

3.4.2 Adaboost makes use of the boosting notion, an ensemble strategy meant to improve weak learners' performance. Using the real data set,it first trains the classifier. Following that, the classifier is trained in several iterations, every attempts correct the mistake of the preceding iteration. Decision tree was used as the default classifier for boosting[27].

3.4.3 Gradient Boosting (GB) Using a serial method, this technique creates an ensemble of trees by training a weak model—one with few splits, for example—and then gradually enhancing its performance by continuing to produce additional trees. Repairing the prior prediction error is the responsibility of each new tree.

3.5. Hyperparameter Optimization

Hyperparameters are crucial, even though they directly affect model performance and the outcomes of training machine learning algorithms[28]. In this work, randomisedsearchcv() is used. RandomisedSearchCV is more efficient than GridSearchCV when processing power or time is limited, as it optimizes hyperparameters by sampling from predefined distributions instead of trying every conceivable combination[29]. The K-fold cross-validation procedure with K = 5 was used to certify the outcome. Using this procedure, 5 groups were formed by dividing the dataset. 4 groups are used for model training, while one group for model testing to assess the model's performance. These model-evaluation approaches are performed 5 times, using different training and test groups each time.

4. Experimental Analysis

This section provides a brief overview of the environment, performance metrics, and classification model results. The experimental setup was conducted on a computer running Windows 11th Core i5-13420H with 16GB of RAM. Python programming language, with all basic libraries: sklearn, matplotlib, pandas, numpy, and seaborn.

4.1 Performance Metrics

Four metric parameters are presented in this work to assess the models' performance. The performance matrices are represented by the following formulas.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

$$Recall = \frac{tp}{tp + fn} \tag{3}$$

$$F1 - score = \frac{2 * precession * recall}{precession + recall} \tag{4}$$

To assess the method's robustness, we first present the confusion matrix for the test data, where tn denotes true negative, tp true positive, fn false negative, and fp false positive. A confusion matrix evaluates the above-mentioned values[30].

4.2 Results and Discussion

This work employed ML techniques, namely RF, Adaboost, and GB, with hyperparameter optimisation, outlier removal, and balancing techniques to predict cardiac disease. From the UCI platform, the authors gathered the Cleveland and statlog datasets, combined them, and used them in their work. 80: 20 ratio for the training parts and testing parts. To detect heart disease, several medical characteristics from the dataset were utilized. Class 0 indicates that a person is disease-free, whereas class 1 suggests that a person has an ailment. These criteria were used to perform classification. F1 score, Accuracy, Precision, and Recall were used to evaluate the model's performance. Before resampling the dataset, it's also important to determine the model's effectiveness on the original, imbalanced test set (before SMOTE) to gain a more realistic understanding of its impact on unseen, real-world data. The metrics for each model are shown on the imbalanced dataset in Table 2.

Table 2. Model performance metrics for imbalanced datasets.

Model	Accuracy	Precesion	Recall	F1 score
Random Forest	97.2	99	94.5	96.7
GB	97	99	94	96
Adaboost	88	90	91	89

From Table 2, it is evident that Random Forest outperformed; however, GB performance is very close to RF. Adaboost lags behind the above two models. Table 3, depicts the model's performance after applying the balancing technique.

Table 3. Performance metrics of models on the balanced dataset.

Model	Accuracy	Precesion	Recall	F1 score
Random Forest	98.2	99.5	96.7	97.9
GB	97.6	99	95.7	97.4
Adaboost	89.3	87.9	91.4	89.4

From Table 3, it is drawn that balancing techniques improve the model's effectiveness. The model on the balanced data showed slightly higher results in all metrics, especially recall and F1-score, which might be preferable depending on the false negative cost in this medical context. Comparison analysis of metrics in a bar graph is shown in Figure 2. Random Forest is the top performer with the highest accuracy of 98.2. The confusion matrix of Random Forest is presented in the Figure. 3.

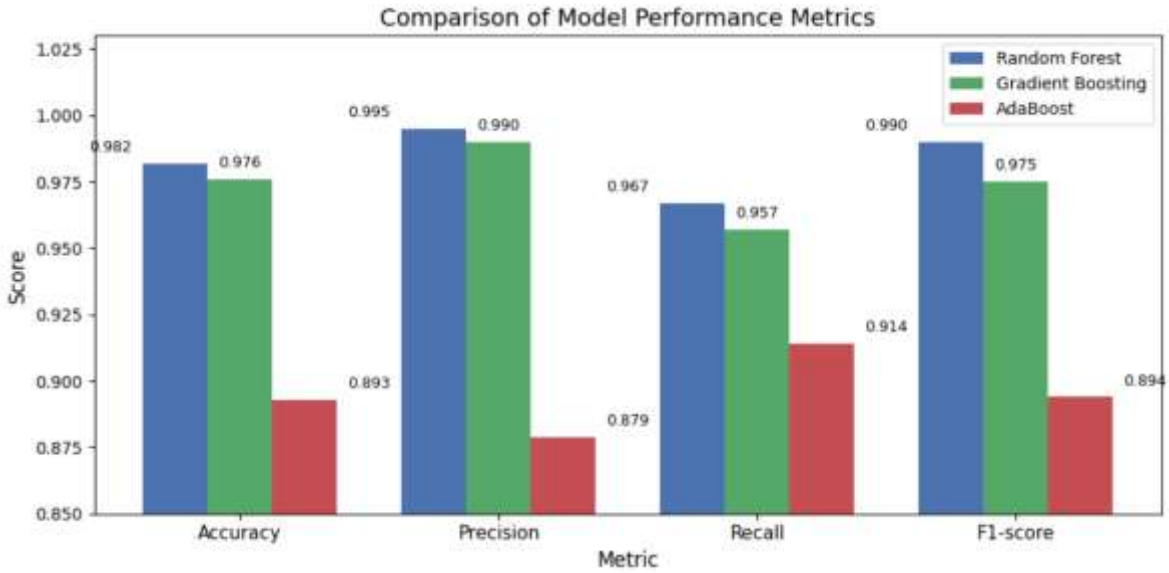

Figure 2: Comparison of metrics

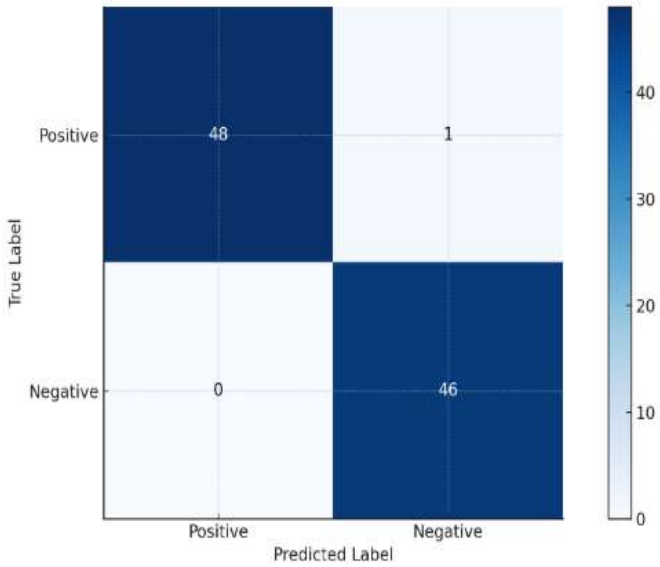

Figure 3: Confusion matrix of Random Forest

5. Conclusions and Future Scope

In summary, our study on CVD prediction using mixed datasets and machine learning approaches has yielded crucial findings with significant ramifications for researchers and healthcare professionals. Our results highlight the inherent difficulties in handling imbalanced datasets and underscore the importance of correctly identifying positive events, especially when their representation is constrained. The efficacy of ML models in forecasting coronary disease has been enhanced by the implementation of balancing techniques and outlier removal from the integrated (Cleveland + Statlog) heart dataset. The five-fold cross-validation approach and the RandomizedSearchCV hyperparameter method have been employed before model implementation to achieve optimal accuracy. Three Machine Learning classifiers were developed and analyzed using F1-score metrics, recall, precision, and accuracy, and RF achieved the highest accuracy of 98.2%, followed by GB 97.6 %.

To demonstrate its potential for healthcare and medical diagnostics, the proposed work will be further evaluated and investigated using a variety of medical datasets in the future. The authors plan to experiment with different outlier identification and balancing strategies in their future work to further improve system performance. Additionally, it is necessary to investigate other hyperparameter optimization methods.

Funding source

No funding was received for this study.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgment

The authors would like to acknowledge DST-FIST (Department of Computer Science & Engineering, Jamia Hamdard) No SR/FST/ET-11/2019/313(C) for providing the facilities to conduct the research.

References

[1] "World Health Organisation," *WHO*, 2020. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] S. Gour, P. Panwar, D. Dwivedi, and C. Mali, "A machine learning approach for heart attack prediction," in *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 1*, 2022, pp. 741–747.

[3] K. Saikumar, V. Rajesh, and B. S. Babu, "Heart disease detection based on feature fusion technique with augmented classification using deep learning technology," *Trait. du Signal*, vol. 39, no. 1, p. 31, 2022.

[4] S. Chopra, P. Agarwal, J. Ahmed, S. S. Biswas, and A. J. Obaid, "Roberta and BERT: Revolutionizing Mental Healthcare Through Natural Language," *SN Comput. Sci.*, vol. 5, no. 7, p. 889, 2024.

[5] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *Appl. Sci.*, vol. 11, no. 18, p. 8352, 2021.

[6] U. Sadar, P. Agarwal, S. Parveen, G. Dhand, and K. Sheoran, "HEART DISEASE PREDICTION USING MACHINE LEARNING CLASSIFIERS WITH VARIOUS BALANCING TECHNIQUES," *Proc. Eng.*, vol. 6, no. 4, pp. 1871–1878, 2024.

[7] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A study of machine learning in healthcare," in *2017 IEEE 41st annual computer software and applications conference (COMPSAC)*, 2017, vol. 2, pp. 236–241.

[8] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mob. Inf. Syst.*, vol. 2020, no. 1, p. 8843115, 2020.

[9] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci. Rep.*, vol. 10, no. 1, p. 19747, 2020.

[10] Sharma, Kavita, Yogita Gigras, Vishnu Sharma, D. Jude Hemanth, and Ramesh Chandra Poonia, eds. Internet of healthcare things: machine learning for security and privacy. John Wiley & Sons, 2022..

[11] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2022, no. 1, p. 1410169, 2022.

[12] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc. Anal.*, vol. 2, p. 100016, 2022.

[13] Uzama Sadar, P. Agarwal, and Suraiya Parveen, "Heart Disease Prediction Using Optimized Feature Selection and Classification Techniques," *CINEFORUM*, vol. 65, no. 3 SE-Journal Article, pp. 529–552, Aug. 2025, [Online]. Available: https://revistadecineforum.com/index.php/cf/article/view/472

[14] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction," *IEEE Access*, vol. 9, pp. 130805–130822, 2021.

[15] J. Ivan and S. Y. Prasetyo, "Heart Disease Prediction Using Ensemble Model and Hyperparameter Optimization," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 8s, pp. 290–295, 2023.

[16] Gupta, R., Gusain, N., Shirole, B. S., Jagtap, M. T., Thomas, S. A., & Kumar, S. A. N. T. O. S. H. (2025). Optimizing healthcare management systems with AI and machine learning. *South Eastern European Journal of Public Health*, 2973-2985.

[17] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data Knowl. Eng.*, vol. 140, p. 102042, 2022.

[18] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Sci. Rep.*, vol. 14, no. 1, p. 23277, 2024.

[19] "Cleveland Heart disease dataset." https://archive.ics.uci.edu/dataset/45/heart+disease

[20] "https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29."

[21] H. M. Zolbanin, D. Delen, and A. H. Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decis. Support Syst.*, vol. 74, pp. 150–161, 2015.

[22] Sapra, P., Paikaray, D., Gusain, N., Abrol, M., Ramesh, S., & Bhardwaj, S. (2023). Evaluation of soft computing in methodology for calculating information protection from parameters of its distribution in social networks: P. Sapra et al. *Soft Computing*, 1-11..

[23] E. M. Abd Allah, D. E. El-Matary, E. M. Eid, and A. S. T. El Dien, "Performance comparison of various machine learning approaches to identify the best one in predicting heart disease," *J. Comput. Commun.*, vol. 10, no. 2, pp. 1–18, 2022.

[24] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in *Information and decision sciences: Proceedings of the 6th international conference on ficta*, 2018, pp. 511–518.

[25] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[26] Gusain, N. (2025). Cardiovascular Disease Prediction through Machine Learning: A Comparative Study of Ensemble Techniques. *Revolutionary Advances in Computing and Electronics: An International Journal*, 27-40.

[27] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, p. 100203, 2019.

[28] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Futur. Gener. Comput. Syst.*, vol.

111, pp. 714–722, 2020.

[29] Y. Rimal, N. Sharma, and A. Alsadoon, “The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms,” *Multimed. Tools Appl.*, vol. 83, no. 30, pp. 74349–74364, 2024.

[30] I. D. Mienye, Y. Sun, and Z. Wang, “Improved sparse autoencoder based artificial neural network approach for prediction of heart disease,” *Informatics Med. Unlocked*, vol. 18, p. 100307, 2020.