iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** X **Month of publication:** October 2025

DOI: https://doi.org/10.22214/ijraset.2025.74614

Hybrid Learning Model for Cardiovascular Risk Assessment

Sonika Harshitha[1], Dr. Sankhya N Nayak[2]
[1]*Student, Department of Computer Science and Engineering, Jawaharlal Nehru new college of engineering,Shivamogga, India*
[2]*Associate. professor, Department of Computer Science and Engineering, Jawaharlal Nehru new college of engineering, Shivamogga, India*

Abstract: Cardiovascular diseases are a global health concern. This study uses a combined machine learning approach to predict heart diseases - it joins Random Forest in addition to Linear Regression. The heart's workings are complex, and this research works to improve prediction accuracy. It does not set up the model structure beforehand. The framework trains on old patient data - this data includes age, cholesterol, blood pressure along with how people live. The Random Forest handles complicated connections, and Linear Regression helps explain them. The study looks how many points must be included - it shows what influences heart disease outcomes. This tells effective knowledge for individual healthcare programs. This hybrid methodology, without reliance on a predetermined model, keeps belief for enhancing quick prediction strategies and contributing to more effective interventions in cardiovascular health.
Keywords: Linear Regression (LR), Random Forest (RF).

I. INTRODUCTION

Project entails a Cardiac Illness Prediction System this is created on a Linear Regression classifier that is fused within a ModelTree ensemble framework. The basic point here is investigation is to form a smart and trustworthy model which apart from the high accuracy in the prediction of heart disease likelihood also can be based on features of clinical and other health indicators of the patient. The dataset was subjected to extensive data preprocessing which included missing value treatment, feature scaling, and feature selection to guarantee the uniformity. The suggested ModelTree-based ensemble not only improves predictive performance but also merges the advantages of the interpretability of linear regression models with the adaptive partitioning of decision trees. The system's effectiveness was thoroughly examined with the use of measures like accuracy, precision, recall, F1-score, and ROC analysis, and was found to be much better than the conventional methods employing a single model such as Random Forest, standalone Linear Regression, and hybrid RF+LR models. This study contributes to the advancement of machine learning applications in medical diagnosis, supporting early detection and timely intervention for patients at risk of cardiac disease.

II. HEART DISEASES PREDICTION APPROACHES

The following are the different works carried out in the given area.

Karna Vishnu Vardhana Reddy et al. [1] it looks forward in enhancing for fast detection of diseases using a Cleveland heart dataset and multiple machine learning classifiers. Among them, Logistic Regression bagging achieved a high ROC score of 0.91, and SMO proved most effective in cardiovascular risk prediction.

M. Snehith Raja et al. [2] used the Random Forest algorithm in a Python-based machine learning system to detect disease early, enabling timely treatment and reducing mortality.

Pronab Ghosh et al. [3] proposed a machine learning model using feature selection techniques for exact cardio disease determination through effective data preprocessing and risk factor identification.

E. I. Elsedimy et al. [4] developed a QPSO-SVM hybrid model combining Quantum-behaved Particle Swarm Optimization with Support Vector Machine to improve heart disease detection accuracy over traditional diagnostic methods.

Nadikatla Chandrashekar and Samineni Peddakrishna et al. [5] employed six ml-based procedures, involving Random Forest and Logistic Regression, using Cleveland and IEEE Dataport datasets, and optimized model performance through GridSearchCV with five-fold cross-validation.

Yong-Suk Jeong et al. [6] analysed heart disease determining by various machine learning algorithms on the NEC-service medical dataset, comprising 4,699 patients aged over 45 diagnosed under the I20–I25 cardiovascular category.

Osman Taylan, et al. [7] highlight wanting to have an accurate CVD detection and advocate using AI tools to predict and classify these complex conditions.
Vijetha Sharma, et al. [8] explore using Decision Trees, for early and support clinical decision-making amid rising cardiovascular risks.
Chintan M Bhatt, et al. [9] demonstrate that machine learning, particularly multilayer perceptron with cross-validation, achieves the highest accuracy in predicting cardiovascular disease using k-modes clustering.
Dhai Eddine Salhi, et al. [10] highlight how machine learning analyses patient dataset to real-time monitoring for improved outcomes.

III. METHODOLOGY

A training set is the data that will be utilized to train a model, teaching it how to identify the data's patterns and relationships; whereas, the testing dataset is to assess the model's performance and consequently, check its generalization ability. The training phase is the stage when the machine-learning algorithm is run on the training set. The algorithm learns from the data, detects regularities, constructs a predictive model which, if it was to analyze future data, could produce very accurate predictions. During the training, it is tested from a test set to check their performance. Metrics like accuracy in contrast the total sum of predictions that are correctly made from model, thus giving a rough idea of its efficacy. If the model is not successful in yielding good outcomes, the reason might be its need for retraining on different data or with an alternative algorithm. Once the model is perceived to be demonstrating the right behavior, its deployment in the factory can be done to forecast on the new data and its usage in the real world.

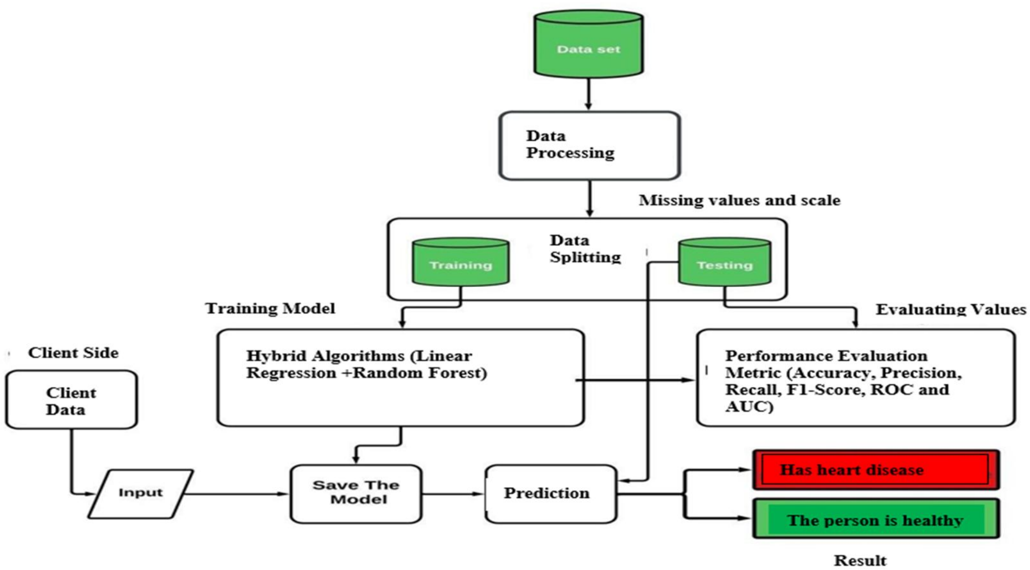


Figure 1: Block diagram for Predicting heart disease.

Figure 1 gives overall workflow of the Cardiac Illness Prediction System. Here's a detailed explanation of each block:

A. Dataset

It has various medical and demographic features such as age, cholesterol level, blood pressure, heart rate, and other health indicators. This dataset forms the foundation for various process.

B. Data Processing

In this stage, cleaning of dataset and preprocessing are performed to prepare it for building of model. This includes:

- Handling values which are missing by SimpleImputer (filling missing entries with mean values).
- Feature scaling using StandardScaler to ensure that all input variables are on the same scale and contribute equally during model training. This step ensures whether it is consistent, reliable, and suitable for other algorithms.

C. Data Splitting

After preprocessing, the dataset is divided into training dataset around 75% and testing dataset around 25%.

D. Training Model (Hybrid Algorithm)

The hybrid algorithm combines Linear Regression (LR) and Random Forest (RF) techniques.

1) Linear Regression values linear relationships which are linear in nature to features and output (disease presence).
2) Random Forest adds robustness by handling non-linear relationships through ensemble learning. Together, they form a hybrid model that balances interpretability and high prediction accuracy.

E. Performance Evaluation

After training, the testing process goes on with metrics like accuracy, Precision, Recall, F1-Score are included. These metrics collectively determine how effective the model is in diagnosing heart disease.

F. Client Side

The client side represents the terminal where patients or healthcare professionals enter values (e.g., age, blood pressure, cholesterol). This allows real-time predictions based on trained models.

G. Save the Model

After training, the model is saved for future use using serialization techniques like joblib or pickle. This eliminates the need to retrain the model every time and enables deployment in web or desktop applications.

H. Prediction

When new data is provided, the saved model processes the input and predicts whether the person has heart disease or not.

I. Result

The system finally displays the result in two possible outputs:

1) Red Box: *"Has heart disease"* → indicates a positive diagnosis.
2) Green Box: *"The person is healthy"* → indicates a negative diagnosis.

IV. IMPLEMENTATION FRAMEWORK

The system's Python version is predominantly rooted in scikit-learn to make its development easy. It is organized into a different module that carry the load in each specific task of the framework like data preprocessing, model building, and evaluation. Main qualities that are managed are data preprocessing, which is the way of dealing with problems such as handling missing data and standardizing features.

Through the utilization of Simple Imputer, which refills the voids, whereas Standard Scaler, which is used to normalize the features, makes them uniform in this way, and thus enhances the model's performance. The most striking feature of the system is the hybrid model, which is the main reason behind the system's prediction abilities., Two algorithms, Random Forest Classifier, and Linear Regression are typically used to create such a model.

The simple relationship between the X (the coefficients related to the risk factors) and Y (the ability to predict the patient's risk of developing complications) variables is represented as: $Y = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + ... + \beta n\ Xn + \varepsilon$. On the contrary, a decision tree classifier that usually recognizes the language independently of Classifier selects trees and outputs it. Reference leaf nodes are those of the heart disease patients as determined by the dataset.

V. RESULT

A user-friendly terminal is capable of processing 14 key information inputs accurately, thus facilitating the rapid and precise results. The charts displayed are the ROC curves for the linear regression, random forest, and hybrid LR-RF model with AUC values being 0.84, 0.85, and 0.98 respectively. The combined model gives result on the position since the upper left corner, with a high TPR and low FPR, is the one nearest to it.

	precision	recall	f1-score	support
(Absence of disease) 0	0.98	0.99	0.98	119
(Presence of disease) 1	0.99	0.98	0.99	138
accuracy			0.98	257
macro avg	0.98	0.98	0.98	257
weighted avg	0.98	0.98	0.98	257

Figure 2: Performance Evaluation Metrics of the Proposed Model

The graphic showing the accuracy, precision, and recall calculations in binary classification situations is stated to be the best model representation by Figure 2. This model shows that the classification problem consists of two contradictory classes which are the target class (positive) and the alternative class (negative).

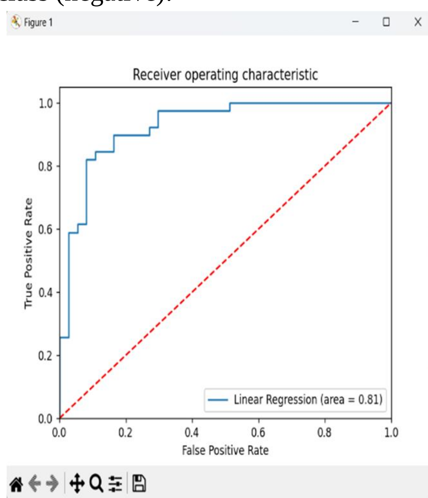


Figure 3: ROC curve for LR

Figure 3, is a depiction of a linear regression classifier's predicting capacity on positive and negative cases. In this instance, the AUC (Area Under the Curve) represents the commendable score of 0.84. The nearby curve to the upper left indicates the model perfect performance.

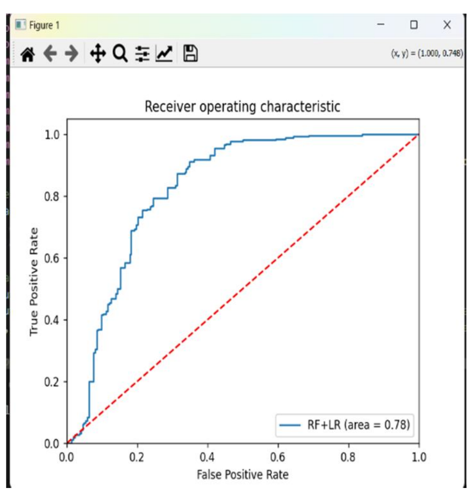


Figure 4: ROC curve for LR-RF

When its accuracy is 0.78, the Hybrid RF and LR model are shown to have the Receiving operating curve (ROC) curve functions displayed in figure 4.

The classification of positive and negative cases by a purely hybrid LR-RF model is presented by figure 5. The second curve indicates an AUC of 0.98 which is equal to or only a little bit higher than the first curve's result. The second model is a little more distant from the first one and closer to the upper left corner, which suggests that the second model is a little bit better.

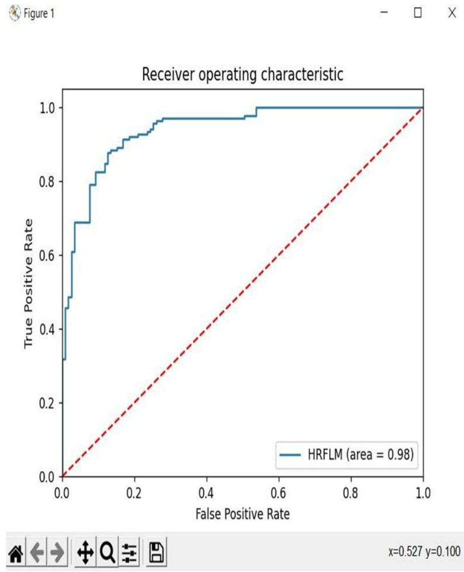


Figure 5: ROC curve for hybrid RF-LR

Table 1: Performance Comparison

SL.NO	MODEL	AUC	DISTANCE TO UPPER LEFT CORNER
01	LR	0.81	0.275
02	RF	0.84	0.25
03	HYBRID MODEL(RF+LR)	0.78	0.34
04	HYBRID MODEL(LR+RF)	0.98	0.02

Table 1 provides a different models performance comparison The measure of distance to the upper left corner is computed using the equation.

$$\text{Distance} = \text{sqrt}\ (FPR * FPR + (1 - TPR) * (1 - TPR))$$

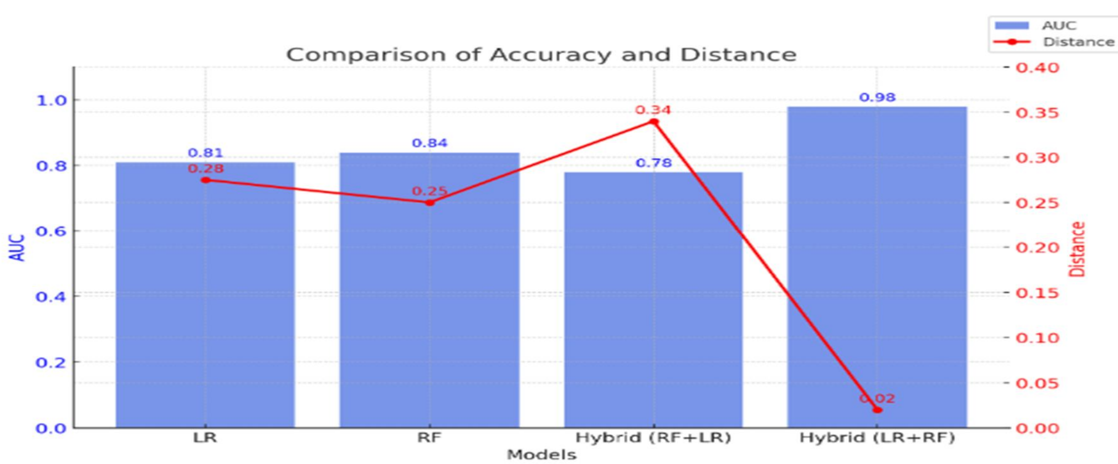


Figure 6: Comparing the accuracy and distance of ML algorithms

The comparative performance of many models for heart disease detection is depicted in the figure 6 through AUC (Area Under the Curve) and Distance, hence the illustrative discussion. The x-axis houses four models that are Hybrid (RF+LR), and Hybrid (LR+RF). The blue bars give the AUC values, which are the measures of success for the model's ability to differentiate healthy patients to the diseased ones while the red line- the distance metric, shows the deviation from the ideal classification performance.

According to the findings, the Hybrid (LR + RF) model emerged as the most dynamic with its AUC of 0.98, which represents the highest score thus far, this is a clear indication of its above-average level of prediction accuracy and robustness. Also, the model achieved the lowest distance value (0.02), which indicates that it is the closest to the ideal state, i.e., a perfect classification boundary. Conversely, the Hybrid (RF + LR) model struggled a lot as it obtained an AUC of only 0.78, by far the worst score and a distance value of 0.34 which clearly shows the low-level performance. The two standalone models Linear Regression (AUC = 0.81, distance = 0.28) and for another algorithm (AUC = 0.84, distance = 0.25) it operates moderately on their own and cannot fulfil the same level of performance as the hybrid one.

The comparison indicates unambiguously that the system could be made more accurate, better balanced, and more reliable in heart disease prediction by employing a hybrid configuration of Linear Regression and Random Forest (LR + RF) in place of the individual or reversed hybrid configurations.

VI. CONCLUSION

Heart diseases is a major challenge, continuous factors and the presence of many risks whose interrelationship is nonlinear. The Random Forest suits well with this issue; the only problem being that it might end up with overfitting, whereas the linear regression might cause failure since it is too simple. So as provided in the text can be the amalgamation of various algorithms. Basic motive to its methodology is to exploit merging of algorithms to improve the rate of accuracy to prediction, give new information about the intricate connections in the heart diseases, and reduce the shortcomings of the individual models. To sum up, this joint method is really prospective for progress in medical technologies and thus improving the diagnosis and therapy of heart diseases will be a step further in fighting the multidimensional problems arising from cardiovascular illnesses.

REFERENCES

[1] 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," *Applied Sciences*, vol. 11, art. 8352, 2021.

[2] M. S. Raja, M. Anurag, C. P. Reddy, and N. Sirisala, "Machine Learning Based Heart Disease Prediction System," in *Proc. Int. Conf. on Computer Communication and Informatics (ICCCI)*, pp. 12–28, 2021.

[3] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. Mehedi Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[4] E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23901–23928, 2023.

[5] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, art. 1210, 2023.

[6] J. O. R. Kim, Y.-S. Jeong, J. H. Kim, J. W. Lee, D. Park, and H.-S. Kim, "Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database," *Diagnostics*, vol. 11, art. 943, 2021.

[7] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," *Biology*, vol. 12, art. 117, 2023

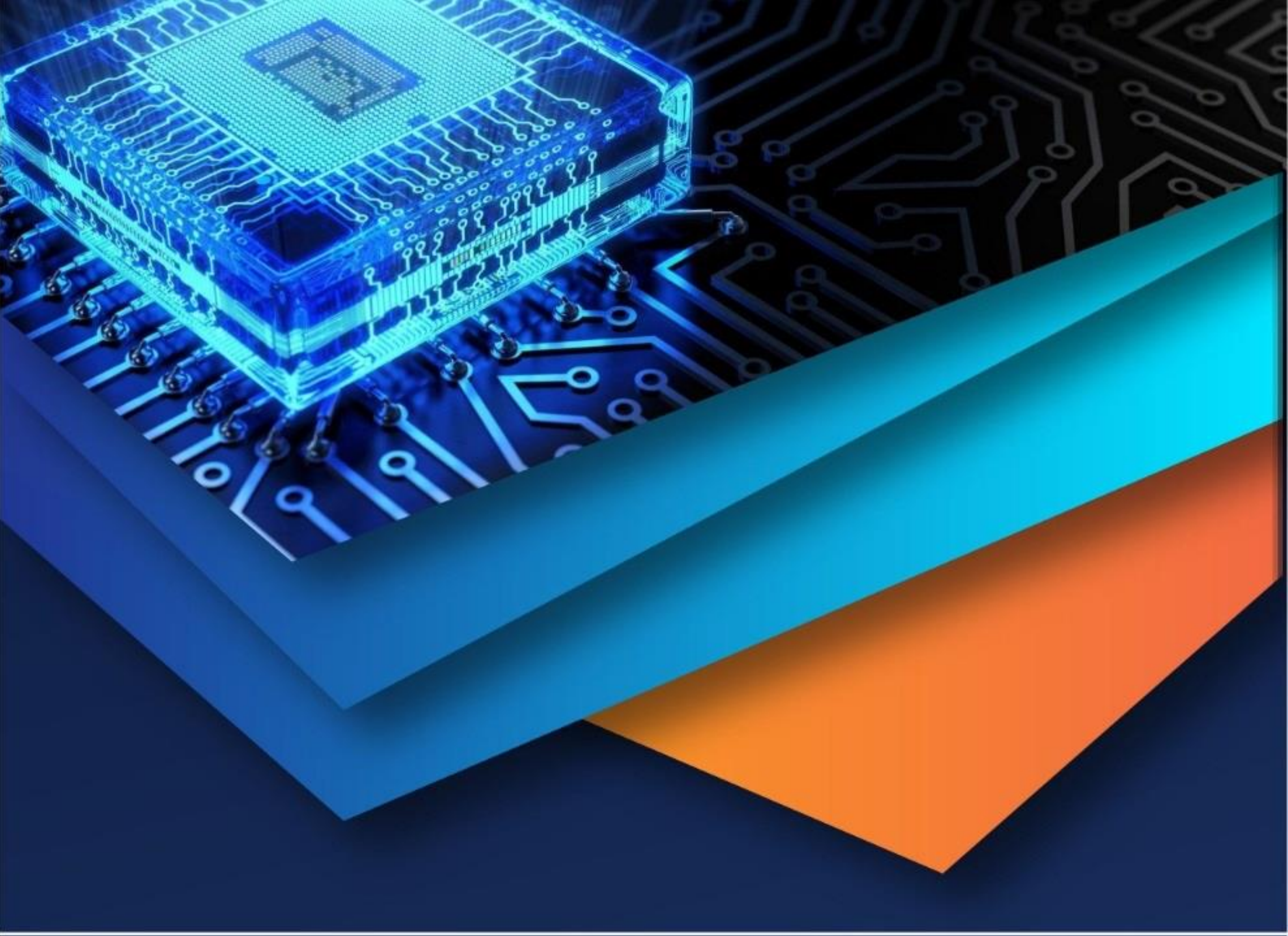
[8] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," in *Proc. 2020 2nd Int. Conf. on Advances in Computing, Communication, Control and Networking (ICACCCN)*, Greater Noida, India, pp. 177–181, Dec. 2020

[9] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, art. 88, 2023

[10] D. E. Salhi, A. Tari, and M.-T. Kechadi, "Using Machine Learning for Heart Disease Prediction," in *Advances in Computing Systems and Applications* (Proc. 4th Conference on Computing Systems and Applications), pp. 70–81, Feb. 2021.

[11] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, Special Issue: Artificial Intelligence Algorithms for Healthcare, vol. 16, no. 2, art. 88, 2023

[12] D. E. Salhi, A. Tari, and M.-T. Kechadi, "Using Machine Learning for Heart Disease Prediction," *Procedia Computer Science*, vol. 199, pp. 628–635, 2021.



10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY