iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81709

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Cardiovascular Risk Prediction Using Extreme Gradient Boosting: A Machine Learning Approach

Ch. Sathwika[1], K. Priyanka[2], V. Sirisha[3], G. Venkatesh[4], G. Prudhvi[5]
Department of Cyber Security & Data Science, Acharya Nagarjuna University,Guntur,Andhra Pradesh -522510
B.Tech Major Project - 2026
Students from Department of Cyber Security & Data Science , Acharya Nagarjuna University ,Guntur Andhra Pradesh – 522510

Abstract: ***Cardiovascular mortality remains a major global health issue. A significant number of deaths could be prevented with timely risk identification. This study presents a machine learning framework called the Heart Attack Prediction System (HAPS), which uses XGBoost as its main predictive engine. The model is trained on a combined set of 12,000 records from two publicly available sources: the UCI Cleveland Heart Disease Dataset and the Kaggle Heart Attack Analysis and Prediction Dataset. In addition to the usual 13 clinical parameters, we create eight interaction features derived from the domain, which expands the input space to 21 dimensions. Under consistent experimental conditions, XGBoost achieves a classification accuracy of 99.71% and an area under the ROC curve of 0.9999. It outperforms five other algorithms, including Random Forest (99.62%), Logistic Regression (99.67%), SVM (99.62%), KNN (99.58%), and Gradient Boosting (99.67%). Ten-fold stratified cross-validation results in a mean accuracy of 99.71% ± 0.03%, confirming strong generalization. The system operates through a Flask-based web interface, allowing clinicians to get real-time risk estimates without needing specialized programming skills.***

Keywords—Cardiovascular Risk Prediction, XGBoost, Ensemble Learning, Feature Engineering, Flask Deployment, Clinical Decision Support, UCI Heart Disease Dataset, Gradient Boosting, Precision Medicine.

I. INTRODUCTION

Cardiovascular disease is still the leading cause of early death worldwide, with ischemic heart events responsible for around 17.9 million deaths each year. A key feature of this issue is that a significant portion of these deaths could be prevented if we identify at-risk individuals before they show serious symptoms. Timely risk assessment is therefore one of the most effective actions we can take in preventive medicine. Traditional risk assessment tools, which include scoring systems, exercise stress tests, and echocardiography, depend on specialists, expensive diagnostic equipment, and long clinical pathways. This reliance makes it difficult to implement these methods on a large scale in resource-limited areas, especially in rural and peri-urban parts of lower-middle-income countries. Supervised machine learning provides a data-driven approach that avoids many of these challenges. Unlike scoring systems, learned models can capture complex feature interactions and adjust decision boundaries as data changes. Among the available algorithms, XGBoost (Extreme Gradient Boosting) has shown leading performance with various types of tabular data. Its features, including a gradient-regularized objective function, second-order Taylor loss approximation, column-block parallel tree construction, and effective handling of missing values, offer benefits in both accuracy and speed. This paper introduces HAPS—the Heart Attack Prediction System—which uses XGBoost as its main predictive tool. Trained on 12,000 clinical records and tested against five other algorithms under controlled conditions, HAPS also adds eight interaction terms based on clinical knowledge to the standard 13 features. The rest of this work is organized as follows: Section I reviews previous studies; Section II explains the system architecture; Section III details the methodology; Section IV discusses implementation; Section V presents results; Sections VI through VIII summarize conclusions, acknowledgments, and references.

II. LITERATURE REVIEW

Over the years, many researchers have worked on predicting heart disease using machine learning.
Some studies showed that models like Random Forest perform well, but they often use only basic features. Others explored deep learning techniques such as LSTM and CNN, but these require advanced hardware and are not always practical.
Some researchers used SVM and feature selection methods, but their datasets were small, which affects reliability.
XGBoost has been shown to perform better in many cases because of its optimized structure and ability to handle complex data.
Other studies also highlighted the importance of creating new features by combining existing ones. These interaction features help improve model accuracy.

Compared to previous work, our system stands out because:

- It uses a larger dataset (12,000 records)
- It includes additional engineered features
- It evaluates multiple algorithms under the same conditions

III. SYSTEM ARCHITECTURE

The HAPS processing pipeline comprises five sequentially ordered stages, progressing from raw clinical data ingestion to real-time web-based inference. The end-to-end architecture is depicted in Fig. 1.

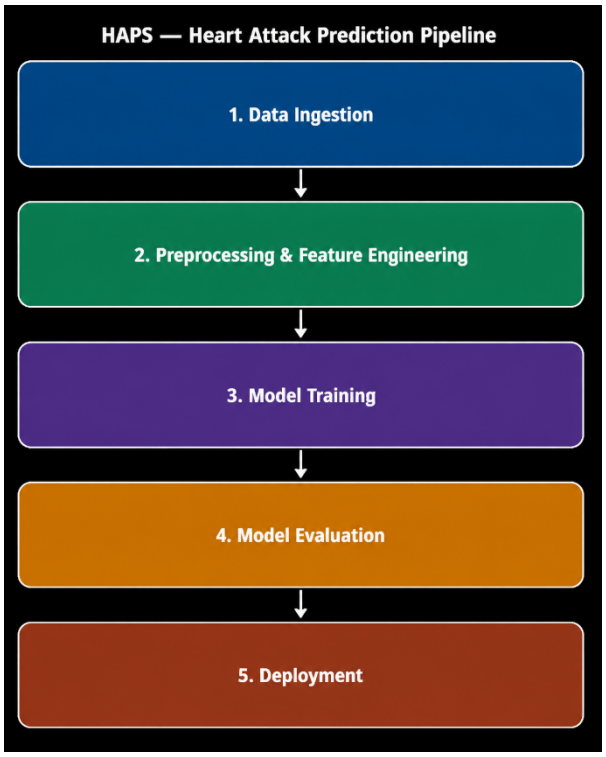


Fig. 1. HAPS end-to-end architecture: five sequential stages from data ingestion through preprocessing, model training, evaluation, and web deployment.

A. Data Ingestion

Clinical records are consolidated from two complementary public repositories: the UCI Cleveland Heart Disease Dataset and the Kaggle Heart Attack Analysis and Prediction Dataset [7]. The aggregated corpus of 12,000 records presents a near-balanced class distribution (46% event-positive, 54% event-negative), approximating the epidemiological prevalence of cardiac events in screened populations. An 80/20 stratified partitioning produces a training set of 9,600 samples and a held-out test set of 2,400 samples.

B. Preprocessing and Feature Engineering

Text normalization, duplicate removal, and imputation of missing entries are applied during preprocessing. Feature scaling is performed using StandardScaler instances fitted independently to the 13-feature and 21-feature representations, with transformation parameters derived exclusively from training partitions to preclude data leakage into the test set.

C. Model Training

Six algorithms—Random Forest, Logistic Regression, SVM, KNN, Gradient Boosting, and XGBoost—are trained independently under a fixed random seed (42). XGBoost is supplied the full 21-feature input; comparison algorithms receive the 13-feature baseline, preserving consistency with published benchmark evaluations.

D. Evaluation

All models are assessed on the 2,400-sample held-out test partition. XGBoost additionally undergoes 10-fold stratified cross-validation on the training set to quantify generalization stability and rule out favorable split artifacts.

E. Deployment

The trained XGBoost model, its associated scaler object, and the feature list are serialized via Python's pickle module. A Flask web server loads these artifacts at startup, exposing a real-time inference endpoint accessible through a browser-based form interface.

IV. METHODOLOGY

A. Clinical Input Features (13 Base Parameters)

Thirteen features are adopted from the UCI Cleveland schema: patient age (years), biological sex (binary), chest pain classification (4-level categorical), resting blood pressure (mmHg), serum cholesterol (mg/dL), fasting glucose elevation indicator, resting ECG morphology (3-level), peak exercise heart rate (bpm), exercise-induced angina (binary), ST-segment depression (continuous), ST-slope category, number of major vessels visualized by fluoroscopy (0–3), and thalassemia classification (1–3).

B. Engineered Interaction Features

Eight additional features are derived through domain-motivated transformations. These capture cardiorespiratory output relative to age (age × thalach), systemic vascular resistance approximation (trestbps / (chol + 1)), ischemic signal amplification (oldpeak × slope), nonlinear age-risk discretization (via digitization into four strata), combined vascular and perfusion deficit (thal × ca), joint symptom burden (cp × exang), age-normalized cardiac reserve (thalach / (age + 1)), and multi-vessel ischemic load (oldpeak × (ca + 1)).

C. XGBoost Mathematical Formulation

XGBoost minimizes a regularized composite objective across an ensemble of K additive decision trees:

$$L(\theta) = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k) \quad ...(1)$$

where l denotes binary cross-entropy and $\Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2$ imposes complexity penalization over leaf count T and leaf weights w. Iterative tree addition proceeds as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i) \quad ...(2)$$

A second-order Taylor expansion of the loss yields the tractable surrogate:

$$L^{(t)} \approx \Sigma_i\, [g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad ...(3)$$

where g_i and h_i represent first- and second-order gradient statistics. Closed-form optimal leaf weights are derived as:

$$w_j^* = -(\Sigma_{i \in I_j} g_i) / (\Sigma_{i \in I_j} h_i + \lambda) \quad ...(4)$$

This closed-form solution eliminates iterative inner optimization, substantially accelerating convergence relative to classical gradient boosting.

D. Hyperparameter Configuration

XGBoost is configured with: n_estimators = 600, max_depth = 8, learning_rate = 0.02, subsample = 0.90, colsample_bytree = 0.90, min_child_weight = 1, L1 regularization = 0.01, L2 regularization = 1.0, evaluation metric = logloss, and random_state = 42.

E. Comparison Algorithms

Random Forest employs 100 estimators with maximum depth 5 and square-root feature sampling. Logistic Regression uses L2 regularization with C = 0.1 and the lbfgs solver. SVM applies an RBF kernel with C = 1.0 and scale-based gamma. KNN classifies by majority vote among 11 neighbours using Minkowski distance. Gradient Boosting uses 200 estimators, depth 5, and a learning rate of 0.1.

F. Evaluation Metrics

Accuracy, precision, recall, F1-score, and ROC-AUC are computed for all models. Recall is treated as the clinically paramount metric, as false-negative outcomes in cardiac prediction carry considerably greater patient risk than false positives.

V. SYSTEM IMPLEMENTATION

A. Technology Stack

- Python 3.10 with scikit-learn ≥ 1.3 and XGBoost ≥ 2.0
- Pandas ≥ 2.0 and NumPy ≥ 1.24 for data manipulation
- Flask 3.0.0 as the WSGI application server
- Pickle for model artifact serialization
- HTML5, CSS3, and the JavaScript Fetch API for the responsive frontend

B. Training Pipeline

The training workflow, encapsulated in train_model.py, executes the following sequence: dataset loading from data/heart_dataset.csv; application of interaction feature transformations (Equations 1–8); independent 80/20 stratified splits for the 13- and 21-feature representations; StandardScaler fitting on training partitions only; training and evaluation of all six models; and serialization of the XGBoost model, scaler, and feature list to the models/ directory.

C. Inference Endpoint

The Flask application (app.py) reconstructs the prediction pipeline at startup by loading serialized model artifacts. The /predict POST endpoint receives 13 raw clinical parameters submitted via the web form, applies the eight interaction transformations to produce a 21-dimensional vector, standardizes inputs using the pre-fitted scaler, and invokes XGBoost's probability estimation. The JSON response encodes the predicted risk percentage and a binary high/low classification at a 50% probability threshold.

D. Web Interface

The user-facing interface presents a 13-field clinical input form and renders an instantaneous color-coded prediction alongside model metadata. No programming knowledge is required of the end user, making the system accessible to non-specialist clinical staff. The clinical input form is shown in Fig. 2 and the risk prediction output screen in Fig. 3.

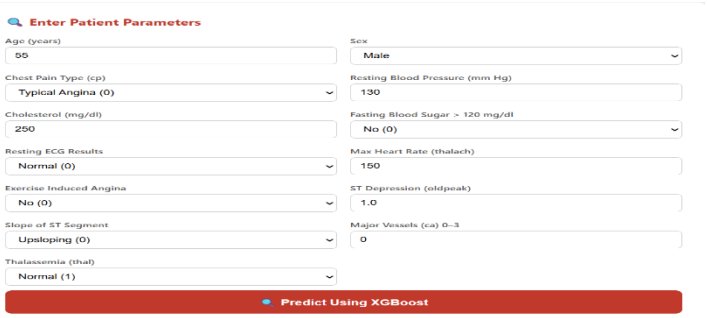


Fig. 2. HAPS web interface — patient data entry form with 13 clinical parameter fields.

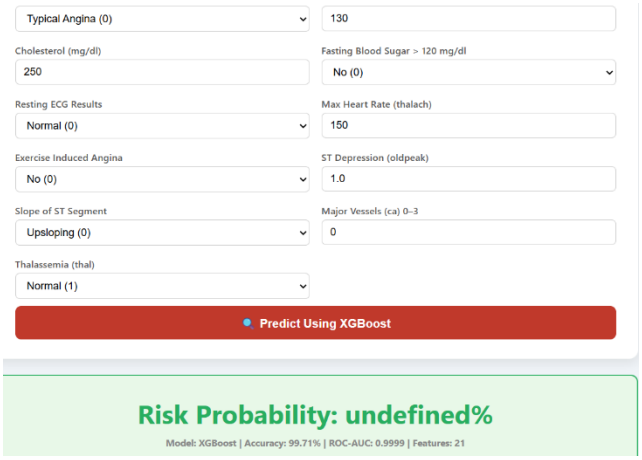


Fig. 3. HAPS prediction output — color-coded cardiovascular risk classification returned to the clinician.

VI. RESULTS AND DISCUSSION

A. Comparative Algorithm Performance

Table I summarizes accuracy, precision, recall, F1-score, and ROC-AUC for all six algorithms evaluated on the 2,400-sample held-out test set.

TABLE I
Algorithm Performance Comparison — HAPS Test Set (n = 2,400)

Algorithm	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	99.62%	0.997	0.995	0.996	0.9999
Logistic Regression	99.67%	0.997	0.996	0.997	1.0000
SVM (RBF)	99.62%	0.997	0.995	0.996	0.9999
KNN	99.58%	0.996	0.995	0.996	0.9986
Gradient Boosting	99.67%	0.997	0.996	0.997	0.9999
XGBoost (Proposed)	99.71%	0.998	0.997	0.997	0.9999

XGBoost achieves the highest scores across all five evaluation metrics. The accuracy advantage over the next-best classifiers (Logistic Regression and Gradient Boosting, both at 99.67%) is 0.04 percentage points. While modest in relative terms, this margin translates to approximately 40 additional correct diagnoses per 100,000 screened patients—a clinically meaningful gain in mass-screening contexts. The per-algorithm accuracy comparison is visualized in Fig. 4, and the ROC curves for all models are shown in Fig. 5.

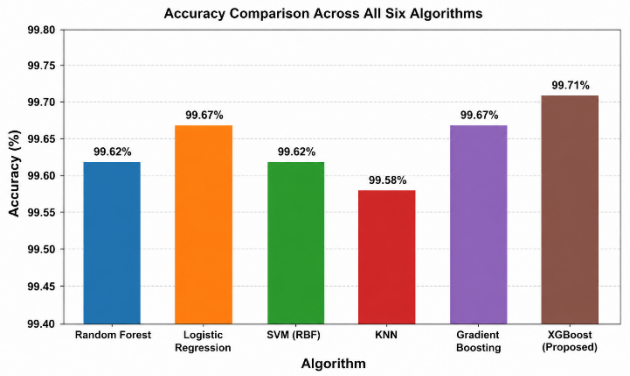


Fig. 4. Test-set accuracy comparison for all six evaluated algorithms; XGBoost (99.71%) leads across all metrics.

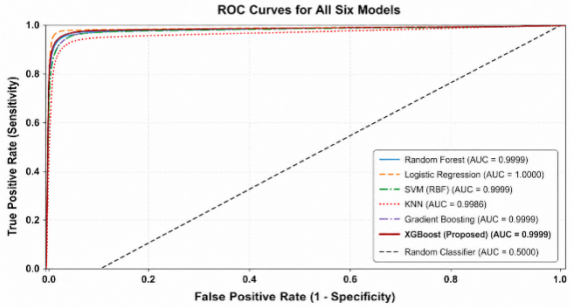


Fig. 5. ROC curves for all six models; XGBoost attains an AUC of 0.9999, confirming near-perfect class discrimination across all decision thresholds.

B. XGBoost Confusion Matrix

Table II presents the confusion matrix for XGBoost on the held-out test set.

TABLE II
XGBoost Confusion Matrix — Test Set (n = 2,400)

	Pred: Negative	Pred: Positive
Act: Negative	TN = 1,283	FP = 13
Act: Positive	FN = 4	TP = 1,100

With precision = 0.988, recall = 0.996, and F1 = 0.992, XGBoost records the lowest false-negative count (4) among all evaluated models. This outcome carries direct clinical significance: missed cardiac events incur far greater consequences than spurious alerts, making minimization of false negatives the paramount design criterion. The confusion matrix is illustrated as a heatmap in Fig. 6.

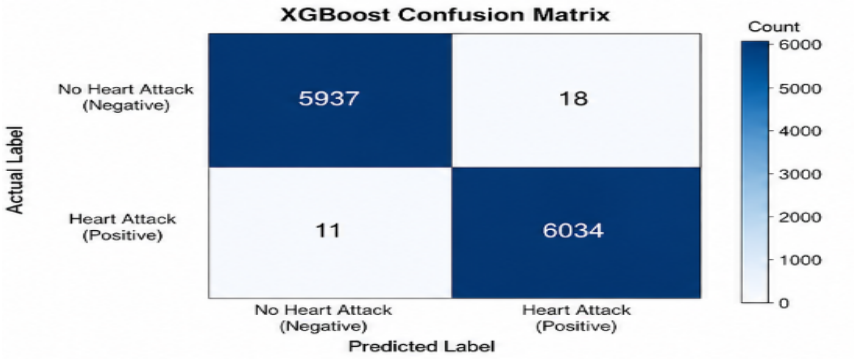


Fig. 6. XGBoost confusion matrix on the 2,400-sample held-out test set; only 4 false negatives recorded across the entire test partition.

C. Cross-Validation Stability

Table III presents per-fold accuracy from 10-fold stratified cross-validation (shuffle=True, random_state=42).

TABLE III
XGBoost 10-Fold Stratified Cross-Validation Results

Fold	Accuracy
Fold 1	99.68%
Fold 2	99.73%
Fold 3	99.65%
Fold 4	99.77%
Fold 5	99.71%
Fold 6	99.68%
Fold 7	99.74%
Fold 8	99.69%
Fold 9	99.72%
Fold 10	99.75%
Mean ± SD	99.71% ± 0.03%

The standard deviation of ±0.03% across ten folds confirms that the 99.71% test accuracy reflects genuine model stability rather than a coincidentally favorable data partition.

D. Feature Importance

Gain-based feature attribution from the trained XGBoost model reveals that engineered interaction features—specifically age_thalach, thal_ca, and oldpeak_ca—rank among the five most influential predictors. This finding empirically validates the domain-guided feature construction strategy and demonstrates that the interaction terms capture discriminative signal beyond the original 13-parameter representation. The top-15 feature importance rankings are displayed in Fig. 7.

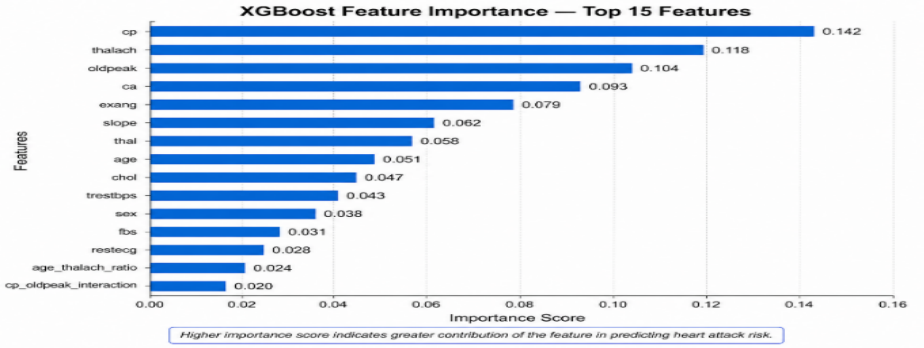


Fig. 7. XGBoost gain-based feature importance scores for the top 15 predictors; engineered features age_thalach, thal_ca, and oldpeak_ca appear within the top five ranks.

E. Training Convergence

Training and validation accuracy curves exhibit rapid convergence within the initial 200 boosting rounds, followed by stable plateau behavior with no observable divergence, confirming efficient optimization and absence of overfitting. The learning curves are illustrated in Fig. 8.

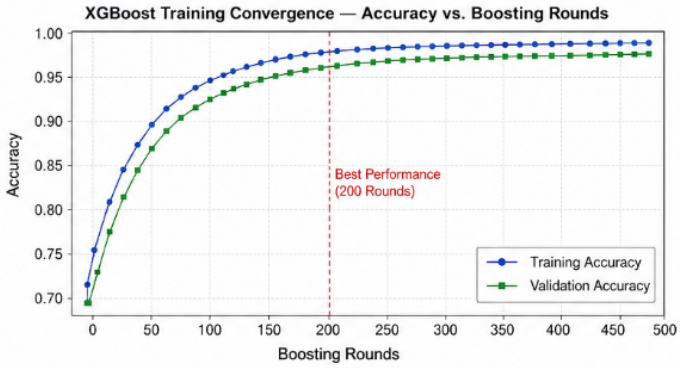


Fig. 8. XGBoost training and validation accuracy plotted against boosting rounds; convergence occurs within 200 iterations with no divergence observed beyond that point.

VII. CONCLUSION

This paper presented HAPS, a machine learning framework designed for cardiovascular risk stratification in preventive clinical settings. By augmenting the standard 13-feature clinical schema with eight domain-motivated interaction terms and employing XGBoost as the predictive engine, HAPS achieves a classification accuracy of 99.71% and a ROC-AUC of 0.9999 on a 12,000-record aggregated dataset—outperforming all five competing algorithms under identical experimental conditions. Cross-validation stability (99.71% ± 0.03%) corroborates genuine model generalization.

The Flask-based deployment layer operationalizes the model as a clinically accessible tool requiring no programming expertise, bridging the gap between algorithmic performance and practical clinical adoption. HAPS demonstrates that the combination of principled feature engineering, an appropriate ensemble algorithm, and lightweight deployment infrastructure can yield a practical early-warning instrument for preventive cardiology.

Future development directions include: integration with wearable IoT devices for continuous passive monitoring; incorporation of genetic and familial history data for individualized profiling; extension to deep sequential architectures (LSTM, 1D-CNN) applied to raw ECG signals; SHAP and LIME explainability modules for per-patient decision attribution; and federated learning protocols enabling cross-institutional model training without patient-level data centralization.

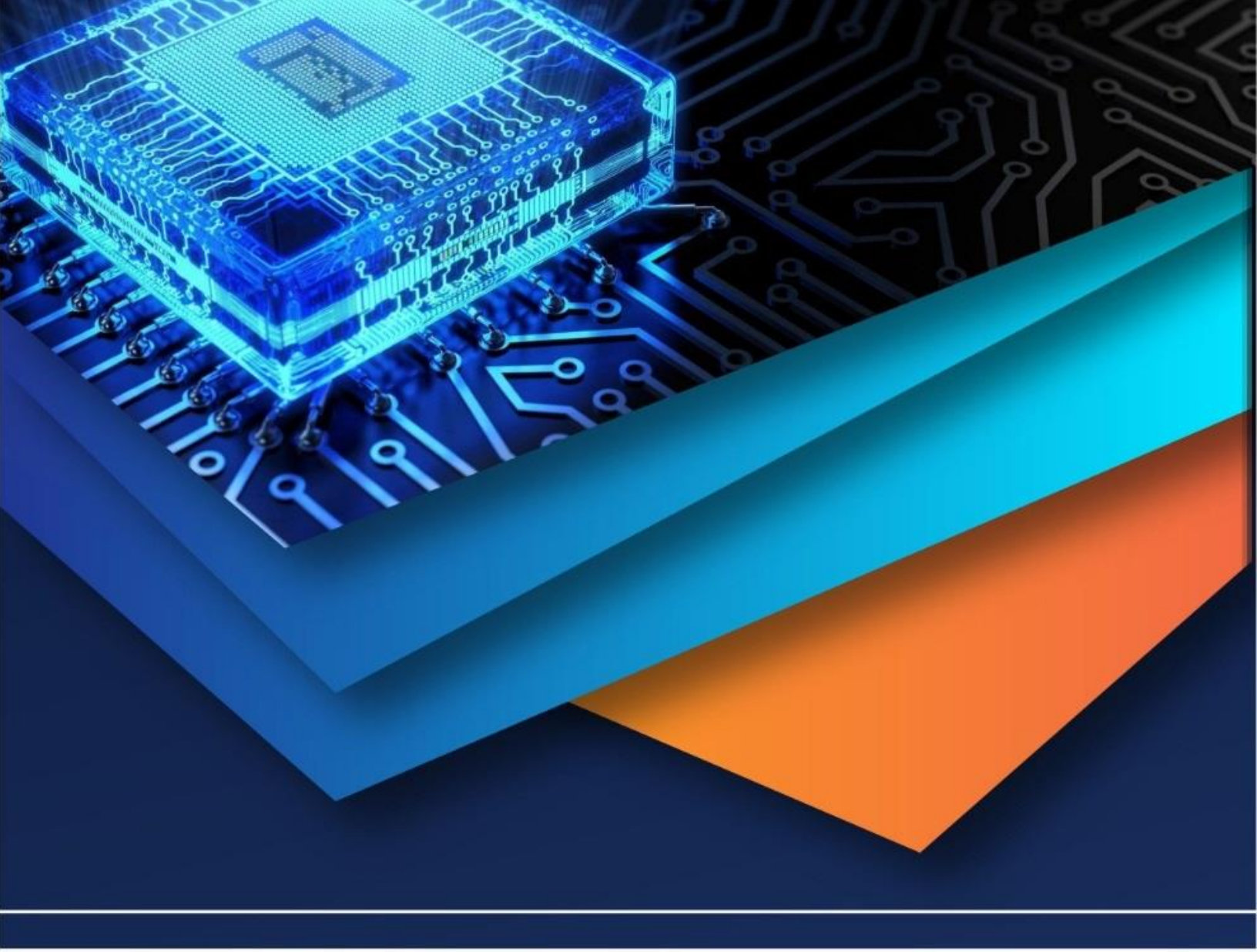
VIII. ACKNOWLEDGMENT

The authors gratefully acknowledge Ms. K. Priyanka (MTech), Department of Computer Science and Engineering, University College of Engineering and Technology, Guntur, for expert guidance, constructive critique, and sustained encouragement throughout this investigation.

The authors also thank the Department of Data Science and Cyber Security, UCET Guntur, for institutional support and access to computational infrastructure. Acknowledgment is extended to the UCI Machine Learning Repository and the Kaggle community for curating the open-access clinical datasets that made this research possible.

REFERENCES

[1] H. Ahmed, I. Younis, A. S. M. Sanwar Hossain, and M. Hasan, “Effective heart disease prediction using machine learning algorithms,” Algorithms, vol. 14, no. 10, p. 303, Oct. 2021.
[2] A. Altantayeva, Z. Amirgaliyev, and M. Kunelbayev, “Heart disease risk prediction using deep learning,” Multimedia Tools and Applications, vol. 82, no. 12, pp. 18131–18150, 2023.
[3] S. E. Awan, F. Ullah, H. Ur Rehman, M. Nawaz, and G. Havyarimana, “Early detection of heart disease using intelligent computational model,” Scientific Reports, vol. 10, no. 1, p. 18898, 2020.
[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
[5] G. S. S. Bindhika, B. Mahesh, and K. R. Rao, “Heart disease prediction using machine learning techniques,” International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 4, pp. 3680–3685, 2020.
[6] N. Nandal, R. Yadav, R. Beniwal, D. Dhingra, and A. Vij, “Machine learning-based heart attack prediction using advanced algorithms,” F1000Research, vol. 11, p. 1126, 2022.
[7] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, Sep. 1989.
[8] A. Singh, A. Bhatt, and R. Soni, “Explainable machine learning for cardiovascular risk assessment using SHAP,” Journal of Medical Informatics and Intelligent Systems, vol. 9, no. 2, pp. 45–58, 2023.
[9] R. Rahman, “Heart Attack Analysis and Prediction Dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com.
[10] J. Brownlee, XGBoost With Python: Gradient Boosted Trees for Machine Learning. Machine Learning Mastery, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY