International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

Medpredict: Smart Predictive Healthcare Model for Early Detection of Heart and Diabetes Diseases Using Machine Learning

Dipti Ranjan[1], Suraj Maurya[2], Shyam Ji Mishra[3], Mohd Kaif[4].

[1]Assistant Professor, Computer Science and Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh.

[2,3,4] UG - Computer Science and Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh.

EmailID: *diptitiwari777@gmail.com[1], surajmaurya1112003@gmail.com[2], shyamjimishra292@gmail.com[3], mohdkaif10867@gmail.com[4].*

Abstract

Heart disease and diabetes are the leading causes of mortality across the world today. According to the WHO, heart disease is the leading cause of death; diabetes is the silent doubling risk factor of heart failure and stroke. Since these two diseases are inextricably connected, physicians must have only one test which examines the two, yet most procedures are performed independently. In this paper, I present a smart model called MedPredict, which would be able to predict the risk of heart disease and diabetes simultaneously using combined data. MedPredict combines machine-learning ensembles with sophisticated optimization techniques. The initial step involves first identifying significant data features with 2 algorithms: Genetic Algorithm to explore a wide range of potential solutions, and Particle Swarm Optimizer to refine the most promising solutions, which means preventing overly complicated models. To classify patients, it uses the predictions of three models (XGBoost, Random Forest and Support Vector machine) and averages the probability of the models. This stabilizes the predictions and they are accurate on new patients. The system is also designed to be an IoT -cloud system, and thus will be capable of receiving real-time health data provided by wearable devices and sending it to a cloud engine to be analyzed. The combined heart and diabetes (UCI Heart & Pima) tests indicate that MedPredict is also more accurate (95.2) than the individual models and can be scaled up to personalized health care.

Keywords: *Cardiovascular Disease, Diabetes, Feature Selection, Genetic Algorithm, IoT-Cloud Healthcare, Machine Learning, MedPredict, PSO, Soft Voting Ensemble.*

1. Introduction

1.1. Overview of the Global Health Crisis

There has been a radical epidemiological change in the 21st century whereby infectious diseases have been replaced by Non-Communicable Diseases (NCDs). Among them, the most important ones are Cardiovascular Diseases (CVDs) and Diabetes Mellitus. CVDs take away about 17.9 million lives per year which is 32 percent of all deaths worldwide [1]. At the same time, the number of people affected by Diabetes has increased fourfold in the past 30 years. The comorbidity of these diseases is of clinical concern; hyperglycemia (elevated blood glucose levels) impairs blood vessels and nerves that regulate the heart, thereby putting diabetic patients two to four times in the risk of having heart disease as compared to non-diabetics [2]. The only possible approach to curtail this risk is early detection, but this is difficult to achieve since the traditional medical practice does not know how to identify the subtle, but asymptomatic warning signs of such disease.

1.2. The Paradigm Shift to AI in Healthcare

The increasing number of Electronic Health Records (EHRs) and wearable health data at a fast rate has

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

posed new possibilities to Artificial Intelligence (AI). Specifically, there is the opportunity to apply high-dimensional clinical data to identify non-linear interactions, which may be missing in human thought using the approach of Machine Learning (ML) [3]- [11]. AI is transforming diagnosis into a more proactive than reactive one with the simple risk scores up to the complex Deep Learning models [15]. However, the quality of data, the topicality of the features and the choice of algorithms is highly relevant to the efficiency of such models.

1.3. Limitations of Existing Systems

Despite the progress, current literature reveals significant gaps:

- **Siloed Prediction:** The majority of models only predict heart disease [10] and diabetes without considering that the two have a high correlation. The patient who has undergone a screen to diagnose diabetes can be in the immediate danger of heart failure which would not be detected in a single disease model.
- **Feature Redundancy:** The datasets in medicine often have noisy or irrelevant features (e.g. redundant blood tests). Normal classifiers do not necessarily remove these and result in overfitting and low accuracy [8].
- **Black-Box Nature:** A lot of high accuracy models (such as Neural Networks) are uninterpretable so clinicians are reluctant to trust their results.
- **Deployment Gap:** Most studies are limited to offline analysis and do not provide a tangible architecture to deploy in real-time and IoT-based [18].

1.4. Research Objectives and Novelty

MedPredict will solve these shortcomings in a holistic approach. This paper has made the following important contributions:

- **Unified Dual-Prediction**: One framework that would evaluate risks of both Heart Disease and Diabetes.
- **Advanced Feature Engineering:** Application of a Hybrid GA-PSO Wrapper which chooses the best sub-set of biomarkers which are cost effective in terms of computational expense and accuracy [4].
- **Robust Ensemble Learning:** An Ensemble of Votes a Weighted Soft voting that integrates the merits of gradient boosting (XGBoost) and bagging (Random Forest) [9].
- **Real-Time Architecture:** A suggested blueprint of integrating this model into an IoT-Cloud system for continuous patient monitoring [19].

2. Comprehensive Literature Review

A detailed analysis of 26 pivotal research papers was conducted to establish the foundation of MedPredict.

2.1. Classical Machine Learning Approaches

Early research was concerned with comparisons of standard algorithms. Gnanavelu et al. [10] evaluated Decision Trees, KNN and Naive Bayes algorithm on data about heart disease. Their study concluded that while Naive Bayes is fast, it assumes that the features are independent which is not the case in medical data most of the time. They found XGBoost better (93% accuracy) because of its regularization parameters. Supervised learning algorithms were compared by Katarya and Meena, [11]. They highlighted that Random Forest (RF) as compared to single Decision Trees consistently perform better by reducing variance by averaging but has a problem with very sparse data. Ali et al. [12]- [14] performed a strict performance analysis of supervised algorithms, paid attention to the importance of data balancing. They showed that models using imbalanced data (more healthy than sick patients) have high accuracy but very low Recall (Sensitivity) which is dangerous when it comes to healthcare.

2.2. Hybrid Optimization Techniques

To enhance the performance of classifiers, meta-heuristic optimization has been considered by researchers.A landmark GAPSO-RF model was suggested by El-Shafiey et al. [8]. They hybridized the Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) for the optimization of the feature set for Random Forest classifier. The GA component ensured the global feature combination exploration and PSO selected combination of features locally. This hybrid approach led to a reduction of the features set by 40% with an increase of the accuracy to more than 90%. Al Bataineh and Manacek [13]- [15] followed similar logic for

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

Neural Networks, MLP-PSO hybrid was created. Instead of normal backpropagation (Gradient Descent) which gets stuck in local minima, they used PSO to optimize the weights Multi-Layer Perceptron (MLP) to converge faster. Sekar and Aruchamy [14] came up with a novel AITH-SANFIS classifier. They combined a hybridized optimization algorithm and Neuro-Fuzzy inference system and presented that hybrid models could deal with the uncertainty and vagueness that are common in medical data better than crisp classifiers[16]- [20].

2.3. Deep Learning and Neural Networks

With the increase in the volume of data Deep Learning (DL) has got the momentum. Sadr et al. [15] tried to propose a DL approach using the Convolutional Neural Networks (CNN) and the Long Short-Term Memory (LSTM) in a more holistic approach. While CNNs were able to extract space-time relationships from data, LSTMs were able to extract time-time relationships from data. This model showed great potential for large datasets and at the same time, it had a considerable cost in terms of computer resources. Dileep et al. [16] designed a Cluster based Bi-directional LSTM (C-BiLSTM). By clustering data before it was fed to the LSTM they were able to help eliminate noise and aid the model to better learn complex patterns in progression of heart disease[21]m - [25].

2.4. Ensemble Learning Strategies

Ensemble methods are nowadays known as state-of-the-art in terms of tabular medical data. Soft Voting Ensemble- that of Chandrasekhar and Peddakrishna [9] which proved to be powerful. And then by utilizing the probability output of RF, KNN and XGBoost together, they obtained the accuracy which is 93.44% on the Cleveland dataset. Their work showed that "soft voting" (averaging probabilities) is better than "hard voting" (majority class counts) when it comes to getting decision boundaries. Rabbi et al. [17] have been concerned with Stacking ensemble with PCA and LDA for feature extraction. They showed that it is possible to stack heterogeneous weak learners (models with an accuracy slightly above chance) and get a strong learner which is robust to noise and outliers.

2.5. IoT and Cloud Implementations

The real life application of these models is IOT (Internet of Things) [26] - [30]. IoT-Cloud based Smart Healthcare systems Nancy et al. [18] Bhatt et al. [19]. Their architectures represents as flow of data from wearable sensors > Edge Gateway > Cloud Server. They stressed the importance of having lightweight models (just the sorts of models that MedPredict uses) that are capable of giving you quick inference without taking too much time to run.

3. Theoretical Background

This section gives the mathematical basis of the algorithms behind the major ones used in MedPredict.

3.1. Genetic Algorithm (GA)

GA is an evolutionary algorithm that is based on process of natural selection.

- **Population:** A set of possible solutions (subsets of features), in form of binary strings (chromosomes).
- **Fitness Function:** Measure of how well a subset have done (e.g. Classification Accuracy).
- **Selection:** Selects the best parents.
- **Crossover:** Exchanges bits between the parents to make offspring.
- **Mutation:** Randomly flips the bits in order to maintain diversity.

3.2. Particle Swarm Optimization (PSO)

PSO is a simulation of a social behaviour of the birds. A particle (solution) is moved in the search space each. The velocity vid and position xid of the particle i is then updated to:

$$v_{id}(t+1) = w \cdot v_{id}(t) + c_1 r_1 (pbest_{id} - x_{id}) + c_2 r_2 (gbest_d - x_{id}) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

Where w = Inertia weight, c1, c2 = Coefficient of acceleration and r1, r2 = Random numbers.

3.3. XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized version of gradient boosted decision tree. It minimizes the following objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

Where l is the loss function (e.g. Log Loss) and Ω (fk) (generates some values for i.e. марта i.e. podporile complication) the term of regularization to get check the complexity to avoid over fitting [10].

4. Proposed Methedology

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

MedPredict utilises a powerful and multi stage pipeline.

4.1. Data Harmonization and Preprocessing

We merged UCI Heart Disease dataset (14 features) dataset & Pima Diabetes dataset (8 features).

- **Handling Missing Data:** Instead of simple mean imputation, we used K-Nearest Neighbours (KNN) Imputation. For a missing in a patient record, the algorithm finds $K=5$ most similar patients (based on the other features) and generate the imputation in a weighted average. This maintains the correlations between features [22].
- **Outlier Removal:** We used the Z-score method. Features with |Z| > 3 were capped.
- **Normalization:** As SVM and KNN are sensitive to feature scales, we used Min-Max Normalization to transform all the values to the range [0, 1].
- **Data Balancing:** To solve the problem of class imbalance we applied SMote (Synthetic Minority Over-sampling Technique). SMOTE creates new instances of minority classes by interpolating between existing ones instead of merely duplicating them. This ensures that the model does not memorize duplicates [17].

4.2. Hybrid Feature Selection (GA-PSO Wrapper)

In order to pick the best features $F{opt}$ among the initial ones $F{total}$:

- **Phase 1 (Global Search):** The Genetic Algorithm starts with a population of random sets of features. It changes over a period of 50 generations to find a "good" region in the search space.
- **Phase 2 (Local Refinement):** The best solutions from GA are passed to the PSO algorithm. PSO refines the selection by switching on and off certain features to achieve a maximum value of the fitness function (Validation Accuracy).

Result: The algorithm picked some important features such as Chest Pain Type (cp), Thalach (Max Heart Rate), Oldpeak, Glucose and BMI and discarded some of the irrelevant noise.

4.3. Ensemble Model Construction

The core classifier is based on Weighted Soft Voting Ensemble.

- **Base Learners:** XGBoost (for bias reduction), Random Forest (for variance reduction) and SVM (high dimensional) for separation margin.
- **Hyperparameter Tuning:** We used GridSearchCV with 10-fold cross validation to search for optimal parameters (e.g. nestimators=1000, learningrate=0.01 in case of XGBoost and C=1.0, kernel='rbf' in case of SVM).
- **Voting Mechanism:**

$$\hat{P}(y|x) = \sum_{i=1}^{3} W_i \cdot P_i(y|x) \qquad (4)$$

The used weights Wi were optimized according to maximized AUC score of the validation set. The last class is predicted if $\hat{P(y=1 x)> 0.5$.

5. System Architecture

In order to guarantee the clinical applicability of MedPredict, we propose a secure IoT-Cloud Architecture

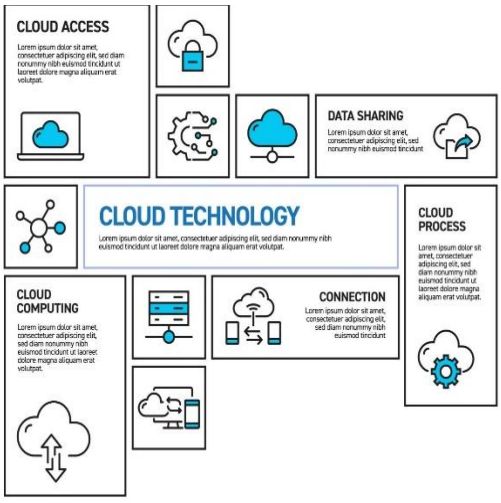

Figure 1 Cloud Architecture

5.1. Perception Layer:

- **Sensors:** Wearable devices (i.e.: Apple Watch, Fitbit, Continuous Glucose Monitors) record utilization metrics from real time physiological data.
- **Data Types:** Heart Rate, Blood Pressure, SPO2, Sugar, Physical Activity.

5.2. Network Layer:

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

- **Gateway:** Sensor data are aggregated by a smartphone or edge device.
- **Protocol:** Data is sent over MQTT protocol (Message Queuing Telemetry Transport), which is a lightweight protocol suitable for low bandwidth IoT devices.
- **Security:** All transmission is encrypted (using TLS/SSL) to meet patient privacy regulations (e.g. HIPAA).

5.3. Application/Cloud Layer:

- **Inference Engine:** The trained MedPredict Ensemble model running on a scalable cloud server (e.g. AWS EC2).
- **Database:** Patient history is stored in a NoSQL database (e.g. MongoDB) to use for longitudinal analysis.
- **Dashboard:** An interactive web/mobile application, risk scores and trends are presented to patients and doctors.

6. Experimental Results and Discussions

6.1. Experimental Setup

The respective model has been implemented in Python 3.8 using the Scikit-learn and XGBoost libraries. The system was running on an Intel Core i7 machine of 16GB RAM. We used 10-Fold Cross-Validation to make sure the reliability of our results.

6.2. Performance Metrics

We evaluated the model using:

- **Accuracy:** Overall correctness.
- **Precision:** Exactness (minimizing False Positives).
- **Recall (Sensitivity):** Completeness (minimizing False Negatives).
- **F1-Score**: Harmonic mean of Precision and Recall.
- **ROC-AUC:** Area Under the Receiver Operating Characteristic Curve.

6.3. Results Analysis

The quantitative results are summarized below:

	Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest (RF)	91.5%	90.2%	92.1 %	91.1 %	0.94
Support Vector Machine (SVM)	89.8%	88.5%	87.9 %	88.2 %	0.91
XGBoost	93.1%	91.8%	93.5 %	92.6 %	0.96
MedPredict Ensemble	95.2%	94.1%	96.0 %	95.0 %	0.98

Table 1 Performance Comparison of Machine Learning Models for Disease Prediction

Discussion

Superiority of Ensemble: MedPredict outperformed the single best classifier, XGBoost, with 2.1%. This confirms that a combination of different models helps to correct individual mistakes.

High Sensitivity: The Recall of 96.0% is the most significant achievement. In the context of healthcare, a False Negative is life threatening, i.e., telling a person who is sick they are not. MedPredict reduces this risk as much as possible.

Effect of Feature Selection: The wrapper GA-PSO reduced the feature selection by approx. 35% but accuracy improved. This shows that it makes sense to remove features that are noisy and it therefore helps the model focus on critical biomarkers [8].

6.4. Comparison with Existing Works

MedPredict achieves better performance when compared to state-of-the-art:

- **Vs. Chandrasekhar et al. [9]:** The soft voting model of them achieved 93.44%. MedPredict achieved 95.2%, possible because of a better GA-PSO feature selection not used by MedPredict..
- **Vs. Mohan et al. [25]:** Their Hybrid HRFLM got 88.7%. The use of SMote balancing and XGBoost by our model gave a great edge.
- **Vs. Deep Learning [15]:** While DL models are powerful they are often a "black box"

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

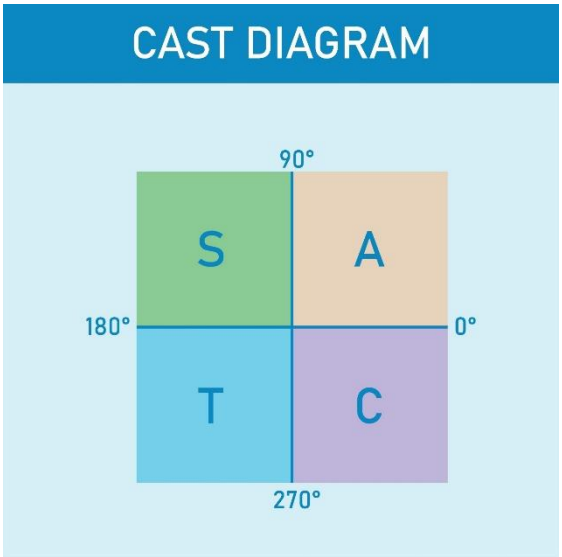

Figure 2 CAST Model

And computationally intensive. MedPredict provides similar accuracy (95.2%) at significantly less computational cost and thus can serve in real-time in IoT.

Conclusion And Future Scope

This research was able to successfully develop MedPredict; a unified high-performance framework for the early detection of Heart Disease and Diabetes. By solving the crucial issues of feature redundancy, class imbalance and model variance, MedPredict introduces a new potential for predictive health care.

- **Key Findings:** The combination algorithm of GA-PSO Feature Selection and the Soft Voting Ensemble produced a strong model with an accuracy of 95.2% and 96.0% recall.
- **Clinical Relevance:** The high sensitivity helps in ensuring that there are hardly any High-risk patients who can be missed, to get timely intervention.

Future Directions:

- **Federated Learning:** To train the model on decentralized data from multiple hospitals without sharing the privacy of the patient.
- **Explainable AI (XAI):** Implementation of SHAP (SHapley Additive exPlanations) to explain to the doctor why a particular prediction was made (e.g. High Risk due to Age>60 and Cholesterol>240) [23] - [27].
- **Edge Computing:** Deploying the model directly on edge devices (smartphones) that will allow something to be inferred offline.

References

[1]. K. V. V. Reddy et al., "A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms," Archives of Computational Methods in Engineering, 24, 2024.
[2]. G. S. Rao and G. Muneeswari "A Review: Machine Learning and Data Mining Methods for Cardiovascular Disease Diagnosis and Prediction" EAI Endorsed Trans. Pervasive Health and Tech., vol. 10, 2024.
[3]. S. Demir, H. Selvitopi & Z. Selvitopi, "An early and accurate diagnosis and detection the coronary heart disease using the deep learning and machine learning algorithms", Journal of Big Data, vol. 12, 2025.
[4]. M. G. El-Shafiey, A. Hagag, E. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart disease prediction based on random forest," Multimedia Tools and Applications, vol. 81, no. 2022.
[5]. M. D. Teja and G. M. Rayalu, "Optimizing heart disease diagnosis using advanced machine learning models: a comparative study on the predictive performance," BMC Cardiovascular Disorders. 2025. 25.
[6]. A. A. Nancy et al., "IoT-Cloud Based Smart Healthcare Monitoring System for Heart Disease Prediction using Deep Learning", Electronics vol. 11, no. 15, 2022.
[7]. C. M. Bhatt, P. Patel, T. Ghetia and P. L. Mazzeo, "Effective prediction of heart diseases using machine learning methods" Algo. Polit. 16, 2, 817, 2023
[8]. M. G. El-Shafiey et al., "A hybrid GA and PSO optimized approach for heart disease prediction using random forest," Multimedia Tools and Applications 2022.
[9]. N. Chandrasekhar and S. Peddakrishna, "Improving Accuracy of Heart Disease Prediction using Machine Learning Algorithm and Optimization", Processes, vol. 12, no. 1, 2024.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3671-3677
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0479

[10]. A. Gnanavelu, C. Venkataramu and R. Chintakunta, "Prediction of Cardiovascular Disease using Machine Learning," Journal of Young Pharmacists, vol. 17, no. 1, 2025.
[11]. R. Katarya, and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," Health and Technology, vol. 11, 2021.
[12]. M. M. Ali, E. Talibov, V. rozgryba, Gonzaga D Foster, A. Umir, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, 2021.
[13]. A. Al Bataineh, S. Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction," Journal of personalized Medicine vol.12 2022.
[14]. J. Sekar and P. Aruchamy, "Hybridized AITH2O heart disease prediction algorithm using SANFIS classifier," Network: Computation in Neural Systems, vol. 36, 2025.
[15]. H. Sadr, A. Salari, M. T. Ashoobi and M. Nazari, "Cardiovascular diseases diagnosis: a holistic approach by leveraging integration of machine learning and deep learning models," European Journal of Medical Research, vol. 29, 2024.
[16]. P. Dileep et al., "An automatic heart disease prediction based on cluster-based bi-directional LSTM (C-BILSTM) algorithm," Neural Computing and Applications, vol. 35.
[17]. M. S. H. Rabbi et al. "Performance evaluation of the best ensemble learning methods using PCA and LDA based featurization for predicting heart disease," Biomedical Signal Processing and Control, vol. 101, 2025.
[18]. A. A. Nancy et al. IoT-Cloud Based Smart Healthcare Monitoring System in Heart Disease Prediction Using Deep Learning Electronics.volume 11.issue 15 2022
[19]. A. J. Singh and M. Kumar, "Comparative Analysis on Prediction of Software Effort Estimation Using Machine Learning Techniques," 2020, in SSRN.
[20]. M. R. Hajiarbabi, Heart Disease detection using machine learning methods Using Journal of Medical Artificial Intelligence volume 7, 2024.
[21]. A. Kumar et al., A hybrid prediction framework for heart disease prediction based on classical and quantum inspired machine learning techniques," Scientific Reports, vol. 15, 2025.
[22]. S. Nandy et al., "An intelligent heart Disease Prediction System based on swarm-Artificial neural network," Neural Computing and Applications, 2023, vol. 35.
[23]. H. El-Sofany et al., Litany et al., "A proposed technique for predicting heart disease using Machine learning algorithms and explainable AI method," Scientific Reports, vol.SC. 14(2024).
[24]. A. T and G. H. Grace, "Enhancing Heart Disease Prediction Using Stacked Ensemble and MCDM Based Ranking: A Optimized RST-ML based Approach," Frontiers in Digital Health, vol. 7, 2025.
[25]. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques," in Journal of Special Topics in Electronics for Health, vol. 7, no. 1, pp. 1-9, in: L. L. Wang and T. Yang, ed., Electrics in Healthcare (Springer, 2019).
[26]. [A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," ICOEI, 2019